Aalto University
School of Science
Master's Degree Program in Computational and Systems Biology

Ronnie Rodrigues Pereira

Identifying Potential *cis*-Regulatory Variants Associated with Allele-Specific Expression

Master's Thesis
Espoo, June 30, 2016

Supervisors: Professor Harri Lähdesmäki, PhD - Aalto University School of Science
Professor Olof Emanuelsson, PhD - Royal Institute of Technology

Aalto University
School of Science
Master's Degree Program in Computational and Systems Biology

ABSTRACT OF THE MASTER'S THESIS

Author: Ronnie Rodrigues Pereira

Title: Identifying Potential cis-Regulatory Variants Associated with Allele-Specific Expression

Number of pages: 75 | Date: 2016/06/30 | Language: English

Professorship: Juho Rousu | Code: T-61

Supervisor at the Home University: Olof Emanuelsson

Supervisor at the Host University: Harri Lähdesmäki

Abstract:

Eukaryotic cellular programs are context dependent. Furthermore, gene regulation exerts this control through layers of interacting molecules. Therefore, genetic variation could interfere in the way these molecules behave by introducing mutations in their underlying DNA sequence. This phenomenon may also selectively affect some gene copies, hence producing an Allele Specific Expression (ASE). This work investigates the effect of genetic variation on genes exhibiting ASE to identify variants with potential regulatory role. Consequently, it shortlists and characterizes variants near protein coding regions as well as introns. Moreover, it relies on analyzing RNA-sequencing and SNP-array data of primary white blood cells of eight European healthy individuals. The study also evaluates two sets of samples: the cells in their basal state and after treatment with lipopolysaccharide (LPS). The proposed model shortlists exon variants representative of ASE, then pairs them against intron/upstream mutations to process linkage disequilibrium analysis. Finally, those which co-segregate with a correlation greater than 0.80 are selected as potential intron/upstream regulatory variants. This process yielded 546 intronic and 80 upstream variants of which 28.20% and 31.25% corresponded to known regulatory elements, according to the Variant Effect Predictor from ENSEMBL. Furthermore, the selected variants are enriched for terms describing an immune response. This trait is especially true for the LPS samples that indeed reacted as if under a bacterial infection. Finally, the selected upstream variants occur more often proximal to the core promoter than to the upper limit of 10kb.

Keywords: Allele Specific Expression, Gene Regulation, Linkage Disequilibrium

ACKNOWLEDGEMENTS

I would like to emphasize my gratitude to Dr. Olof Emanuelsson for supervising this work. His guidance and support have greatly contributed to the outcome of this project.

I would also like to recognize the support of the Education, Audiovisual and Culture Executive Agency – EACEA of the European Union for providing me with the financial means to pursue a master degree in computational and systems biology. The Erasmus Mundus experience greatly expanded my scientific knowledge and cultural worldview.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Eukaryotic cells operate through layers of gene regulation to orchestrate the expression of diverse cellular programs. In addition, genetic variation interferes with these mechanisms by introducing mutations to the genetic information encoded in the DNA molecule. Consequently, variants in regulatory regions may affect gene expression. In fact, deleterious genetic variation may lead to disruptive processes and the possibility of developing complex diseases (Gaffney, 2013).

These traits have been extensively investigated by Genome Wide Association Studies (GWAS). This research area dedicates much effort to identify and quantify genetic loci (eQTL) that are representative of specific phenotypes (Albert & Kruglyak, 2015). In addition, the standard method for evaluating eQTL consists of a two-step strategy: data normalization across samples and linear regression. This procedure uses a linear model to assess the importance of the *loci* to the emergence of the phenotype (Sun, 2012).

Another strategy for eQTL evaluation takes in consideration the phenomenon of Allele Specific Expresion. This behavior corresponds to an imbalance in the expression profile arising from the different copies of a gene (Alberts et al., 2014). Therefore, WASP (van de Geijn, McVicker, Gilad, & Pritchard, 2015) extracts read count information from phased samples and shortlists variants with some measure of association to the phenotype under analysis. This approach does not perform normalization on the data from different samples, but relies on phased genotypes. Moreover, it does not provide information on the regulatory potential of the mutations.

Genetic variation might be the cause of the imbalance in gene expression from ASE genes. Therefore, (Lefebvre et al., 2012) argues for the analysis of linkage disequilibrium between homozygous and heterozygous individuals. This strategy selects variants with potential regulatory role by comparing the rate of inheritance between variant pairs. Thus, mutations segregating together and in short genomic distance are predisposed to contribute towards the analyzed phenotype. In addition, this method heavily relies on identifying candidate pairs that are representative of the phenotype, hence many to many comparisons could quickly arise.

This study investigates the role of base substitutions in the neighborhood region of genes exhibiting ASE. Therefore, it implements quantile normalization to compare the expression profile across different samples from white blood cells at their basal state or after the stimulation of an immune response. Moreover, this work benefits from the gene selection performed by the GeneiASE software (Edsgärd et al., 2016), since it evaluates the potential of a gene to exhibit ASE directly from read counts without the need of phased samples.

This work also proposes an approach to determine the significance of the *loci* to the trait. Consequently, it models the normalized data according to pre-specified haplotype plot patterns and not based on linear regression.

This operation produces a list of variants that serve as anchors for pairing against intronic and nearby mutations. Then, these variant pairs are subjected to linkage disequilibrium analysis. The pairs

with high correlation for co-segregation contain the variants with potential regulatory role. Therefore, the evaluation of the method relies on their functional annotation.

In general terms the research question of this work could be phrased as follows: *How are the genetic variants of ASE genes in primary white blood cells of eight European individuals distributed and which biological processes they affect?* Furthermore, the goals of the project include the following specific objectives.

a) Investigate selection criteria to shortlist ASE genes;
b) Evaluate regulatory region potential by characterizing the variants according to their genomic position, linkage disequilibrium and functional role;
c) Determine the effect of an inflammatory response on the number and identity of the genes;
d) Perform functional enrichment analysis on the affected genes.

In summary, the thesis presents a review of the literature on the topics concerning gene regulation, genetic inheritance and functional enrichment analysis. Then, it describes the stages of the work flow implemented in this study as well as the model assumptions and mathematical framework. This section is followed by the results and their discussion before arriving at some concluding remarks.

2 LITERATURE REVIEW

2.1 Genetic Variation

The structural composition of genes remained unresolved for the greater part of the XX century. This problem generated a substantial debate over the years, thus the understanding of the flow of information has greatly evolved (Gregory, 2005). Researchers of the first half of the last century proposed that genes were part of an intricate system responsible for the emergence of the physical traits in an organism. Furthermore, they believed that either the genes were chemical entities in a catalytic reaction or they controlled the specificity of such enzymes (Beadle & Tatum, 1941).

In 1941, Beadle and Tatum showed that genes regulated the expression of growth factors. This study involved assessing the role of genes by inducing mutations in *Neurospora* through x-ray treatment (Beadle & Tatum, 1941). In addition, the authors examined the impact of the mutations in complex traits influencing development and function. Their findings proved that regulation indeed occurs at the gene level and that it goes beyond superficial phenotypic characters, since this strategy altered pathways that are essential to the survival of the organism. These complex traits are often associated with more than one level of gene regulation (Mannervik, Nibu, Zhang, & Levine, 1999), hence, unlike some of their predecessors, the authors supported the belief of a network of interactions. This understanding was rudimental, because it was not able to describe the intermediate reactions that led to the phenotypic alteration.

In 1944, Avery, MacLeod and McCarty fractionated the lethal pneumonia bacteria (*Streptococcus pneumoniae*) into its constituent parts (DNA, RNA, proteins, lipid and carbohydrates). Later, the researchers individually transformed each part into the harmless pneumonia bacterial strain. Consequently, only the bacteria containing DNA turned into the lethal strain. This experiment proved that the DNA molecule carried the heritable material (Alberts et al., 2014).

Another important step towards describing the identity of genes constitutes the study of the genome size. This quantity illustrates that the total DNA amount in a cell remains constant across most cell types within an organism. Consequently, by 1950, there was a growing belief that genes are not enzymes but DNA (Gregory, 2005).

In 1958, Francis Crick proposed the central dogma of molecular biology (Crick, 1970). This concept stipulated the fundamental actors in the process by which genetic information flows in the cell (Figure 2.1). Crick restated the dogma in 1970 and categorized the reactions into three types of transfer: general, special and unknown (Crick, 1970). The latter are reactions that could never occur while the special transfer may rarely develop. Nonetheless, the general transfer is ubiquitous. It postulates that DNA may be replicated, which yields another DNA molecule, or transcribed, resulting in an RNA molecule. Furthermore, RNA acts as template for proteins in a process called translation.

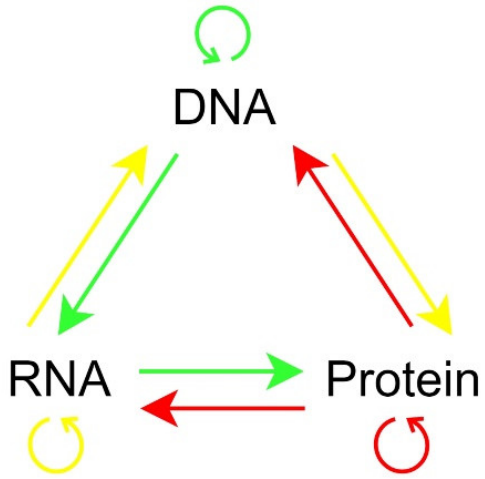


Figure 2.1 – The transfers involved in the flow of genetic information as proposed by Francis Crick. (Adapted from (Crick, 1970))

Green = General; Yellow = Special; and Red = Unknown.

The findings of Francis Crick enabled a fuller understanding of the process involved in protein expression. These discoveries led Jacob and Monod to propose a mechanistic model of gene regulation to explain gene expression in bacteria (Jacob & Monod, 1961). This seminal work strongly contributed to the current understanding of gene regulation. The authors advocate that some molecules will bind to places close to protein coding regions on the DNA, hence the interaction between these molecules and the transcriptional machinery dictates the protein expression.

This notion undeniably places genes in the context of the genetic information encoded by the DNA molecule. However, it raised concerns on the description of organismal complexity, since the genome size does not correlate with gene number in eukaryotes (Gregory, 2005; Levine & Tijan, 2003). For instance, Figure 2.2 illustrates the main components of the human genome. This type of DNA molecule comprises approximately 1.5% of protein coding regions (20,000 - 25,000 proteins), 26% of intronic regions, and 45% of transposable elements. Therefore, most of the genetic information lies in non-coding regions.

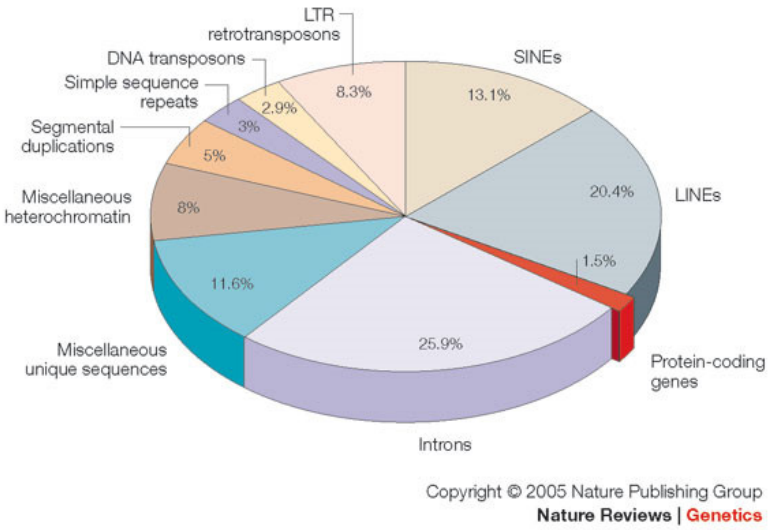


Figure 2.2 – Genome components of eukaryotic organisms. (Gregory, 2005)

Furthermore, the number of genes alone does not reflect organismal complexity. This concept becomes evident in the comparison between the genome of *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. The genome of the flat worm contains approximately 20,000 genes, however, it does not have as many cell and tissue types as the fruit fly, with less than 14,000 genes. In contrast, humans have less than 30,000 genes and exhibit much more complex behavioral and physiological systems (Levine & Tijan, 2003).

This apparent dichotomy may be explained by the different gene regulatory organization of each organism. In fact, two major types of complexes coordinate gene expression: the transcription initiation complex; and, the chromatin modifying complex. These structures act as expression switches, since the regulators alter the binding affinity of the RNA polymerase II to the promoter region (Levine & Tijan, 2003) while the chromatin state dictates the accessibility of this region (Jaenisch & Bird, 2003).

Figure 2.3a displays a simplified schematic of a gene cassette. This drawing portrays the promoter region immediately before the Transcription Start Site (TSS, right arrow). This region harbors many, but not all, of the regulatory elements. Furthermore, it can span from 500 bp (Levine & Tijan,

2003) until 10,000 or 100,000 bp (Alberts et al., 2014; Levine & Tijan, 2003). This range requires a terminology to address the elements influencing transcription. Therefore, proximal sequences lie close to the TSS while distal elements remain upstream of this area. Finally, regulatory elements may also appear in non-coding regions downstream of the TSS (Figure 2.3b), due to the presence of introns in the composition of the protein coding regions.

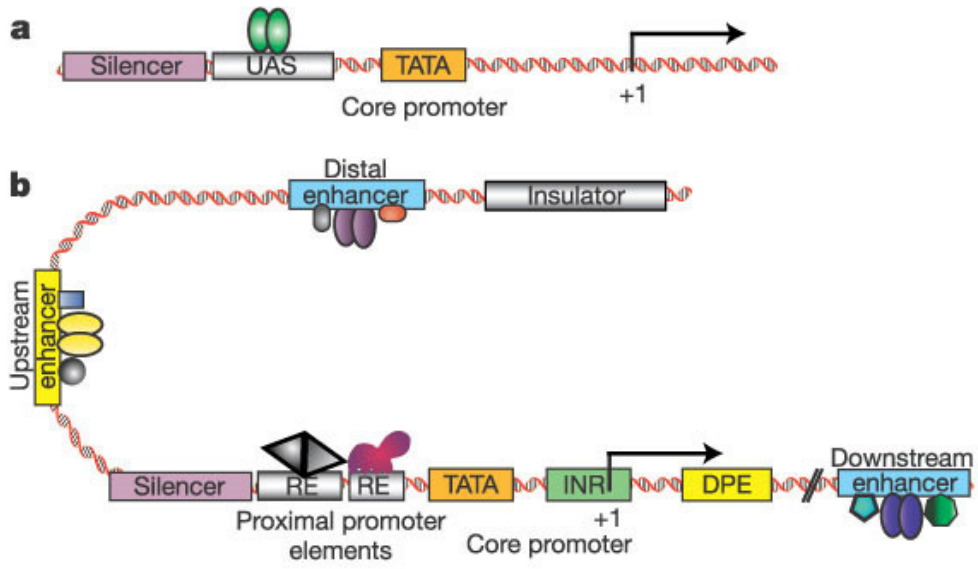


Figure 2.3 – Eukaryotic Transcriptional Units (Orientation: 5' – 3'). a) general yeast gene cassette; b) complex mammalian modules. (Levine & Tijan, 2003)

Note: TATA – Core Promoter
UAS – Upstream Activator Sequence
INR – Initiator Sequences
DPE – Downstream Promoter Elements
RE – Regulatory Elements

The formation of the transcription initiation complex depends on transcription regulators. For example, the TATA Binding Protein (TBP) consists of one of these proteins. Furthermore, TBP belongs to a class of proteins denoted by the name of General Transcription Factors (GTF) (Maston, G. A. et al., 2006). This protein recognizes the TATA box sequence in the promoter region and binds to the DNA double helix. This action leads to a conformational change of the DNA molecule. This entity bends allowing the binding of more transcription factors that ultimately position the RNA polymerase II in the exposed template strand of DNA, thus encouraging the beginning of transcription (Alberts et al., 2014).

This type of assembly procedure leads to low protein production. In fact, it consists of the basal transcription (Maston, G. A. et al., 2006) or the leakiness effect in synthetic biology (Alberts et al., 2014). The formation of the transcription initiation complex (TIC) depends not only on the core promoter but also on the regions located around it. Therefore, the multiple binding regions confer a combinatorial characteristic to the regulation of gene expression (Maston, G. A. et al., 2006). This trait

implies that a small number of regulator proteins may decide on the transcription fate by means of intricate calculations. In other words, the level of protein production depends on the interactions between the TIC, transcription factors and the DNA binding regions.

The *loci* located towards the 5' end of the promoter contain sequences that allow the binding of two types of transcription factors. The first enhances the transcription rate by increasing the binding affinity of the region to the TBP while the second silences mRNA production by repressing the assembly of the transcription initiation complex (Alberts et al., 2014). In addition, complex organisms may exhibit these structures also in the region towards the 3' end of the regulatory region (Levine & Tijan, 2003), as depicted in Figure 2.3b.

Enhancers and silencers are sequence specific modules that present many docking places for transcription factors. Each module generally works regardless of their orientation or distance to the promoter. However, the sequence of binding docks within a module does depend on the orientation (Maston, G. A. et al., 2006). Furthermore, their respective transcription factors often work cooperatively (Alon, 2007), thus a mutation on the binding sequence of any such element decreases the binding affinity of subsequent factors which decreases transcription.

Figure 2.3b also shows the proximal promoter elements. Their function is to enhance transcription by serving as the binding dock to activator transcription factors (Levine & Tijan, 2003). Consequently, this region works similar to an enhancer module.

Another important element for transcription is the core promoter region. This area dictates the direction of transcription and harbors sequence specific binding motifs, such as the TATA box (Levine & Tijan, 2003). (Gershenzon & Ioshikhes, 2005) analyzed the composition of predicted promoter regions and they argue that there are four main types of motifs: TATA, Inr, DPE and BRE. Furthermore, in the same analysis, the TATA box only occurs in one eighth of the results while the more extreme scenario is when the gene carries unknown or unusual motifs. Therefore, the structural variation of the core promoter affects the assembly of the transcription initiation complex because of the varied levels of affinity exerted by the different motifs.

Figure 2.4 illustrates the role of enhancers, silencers, insulators and locus control regions on transcription. Enhancers lead to activation of transcription while silencers work towards repression. Moreover, insulators dictate the gene boundary by preventing run-on transcription. These elements may also physically separate the influence of a repressive chromatin, since they span 0.5 – 3 kb (Maston, G. A. et al., 2006). In addition, it often occurs in areas that are rich in coding or regulatory regions (Fourel, Magdinier, & Gilson, 2004). Finally, locus control regions are responsible for the regulation of clusters of genes. They harbor enhancers, silencers, insulators modules as well as docking sequences for chromatin binding proteins (Maston, G. A. et al., 2006). The resulting cis-regulation often leads to transcription activation. However, this outcome depends on the orientation or module sequence inside the locus control region.

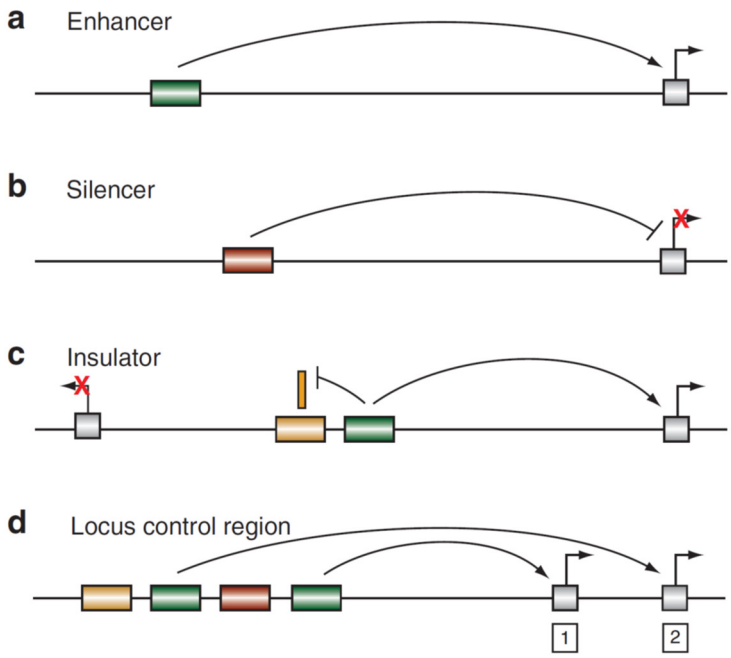


Figure 2.4 – Common regulatory regions and their effect on transcription. (Maston, G. A. et al., 2006)

These regulatory elements all rely on the effective binding of proteins to sequence specific regions on the DNA molecule. Therefore, the transcription rate of a given gene is modulated by the transcription factors and the DNA sequences of the regulatory modules both proximal and distal to the gene (Alberts et al., 2014). These layers of control allow the temporal and spatial behavior of gene expression.

The mutation of such regulatory sequences may also affect transcription rates. These variants may occur due to failure in the proof reading mechanism of the DNA replication (Alon, 2007) or more generally in its repair mechanisms (Alberts et al., 2014). The presence of such variants in germ cells causes the progeny of an individual to also carry the same mutation. In contrast, nucleotide changes to somatic cells occur during the lifespan of an individual and have no hereditary consequences.

In fact, variants conferring stronger binding efficiency in enhancer modules will likely increase gene expression while a deleterious mutation will work to the opposite effect. Similarly, a mutation in a silencer region that elicits higher affinity for a repressive transcription factor leads to decreased transcription. This outcome is the corollary of the interplay between the bound/unbound states of transcription factors and the difference in time scales between their binding to the DNA molecule and the initiation of the transcription process (Alon, 2007).

The importance of categorizing these elements as well as gene boundaries, functional roles and chromatin accessibility produced much discussion in the scientific community over the years. Consequently, there are specialized databases that store this type of information and provide reliable annotation. Some examples of such databases include the ENCODE project (The ENCODE Project Consortium et al., 2012), ENSEMBL (Yates et al., 2016), and TFCat (Fulton et al., 2009). The first two

consists of the international collaboration between research labs to supply data on varied functional elements while TFCat is a curated database dedicated exclusively to transcription factors.

Gene expression is a highly regulated process. It involves many chemical species and interactions between these proteins and the DNA molecule. In addition, mutations may disrupt the optimal folding of transcription factors as well as their binding affinity to regulatory regions. Consequently, these events influence transcription rates. In summary, this section provided an overview of the key players and the interactions involved in the control of gene expression while the next will explore in more detail the behavior of inheritance and gene expression.

2.1.1 Genetic Variation in Humans

The central dogma of molecular biology has established that genes are nucleotide sequences imprinted on the DNA molecule. Furthermore, the current discussion described the processes involved in gene expression by means of its regulatory control. However, it is important to explore the processes related to the inheritance of this genetic material, since these mechanisms explain the genetic variation present in individuals of a population.

Figure 2.5 depicts the chromosome as the most compact form to store the genetic information of an individual. The number of chromosomes employed on this task depends more on the identity of the organism than on its size (Levine & Tijan, 2003). In addition, individuals that reproduce sexually inherit one set of chromosomes from each parent. This set contains one full copy of the genes required for an organism to function properly. The collection of chromosomes in a set forms a haplotype while the collection of sets constitutes a diploid cell. Most of the cells in an organism are diploid with the exception of the gametes. The latter are the cells responsible for fertilization, hence the formation of progeny (Alberts et al., 2014).

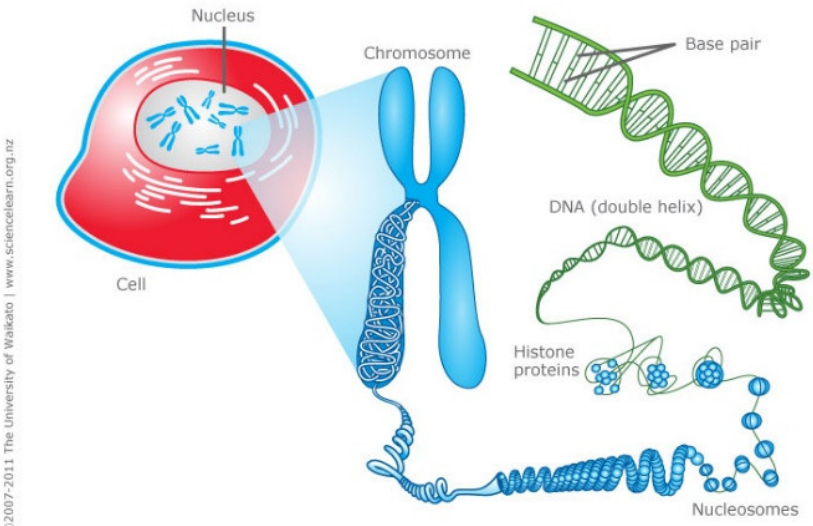


Figure 2.5 – Organization of the genetic information in a eukaryotic cell. (University of Waikato, 2011)

Diploid cells must segregate its haplotypes to form gametes in a process called meiosis. Figure 2.6a shows the steps involved in this mechanism, namely: DNA duplication; homolog pairing; recombination; chromosome segregation and cell division. This procedure entails the production of four gamete cells for each diploid cell. Therefore, the initial step is the duplication of the number of chromosomes, followed by pairing each chromosome copy (maternal-paternal). The parental homologs have their own individual haplotypes, but the recombination step mixes this information by overlapping and swapping stretches of chromosomes from the individuals. This process is the major contributor to genetic variation, since there is an exchange of physical parts between the two homologs. The final steps consist of aligning the chromosomes and pulling them apart into new cells (Alberts et al., 2014).

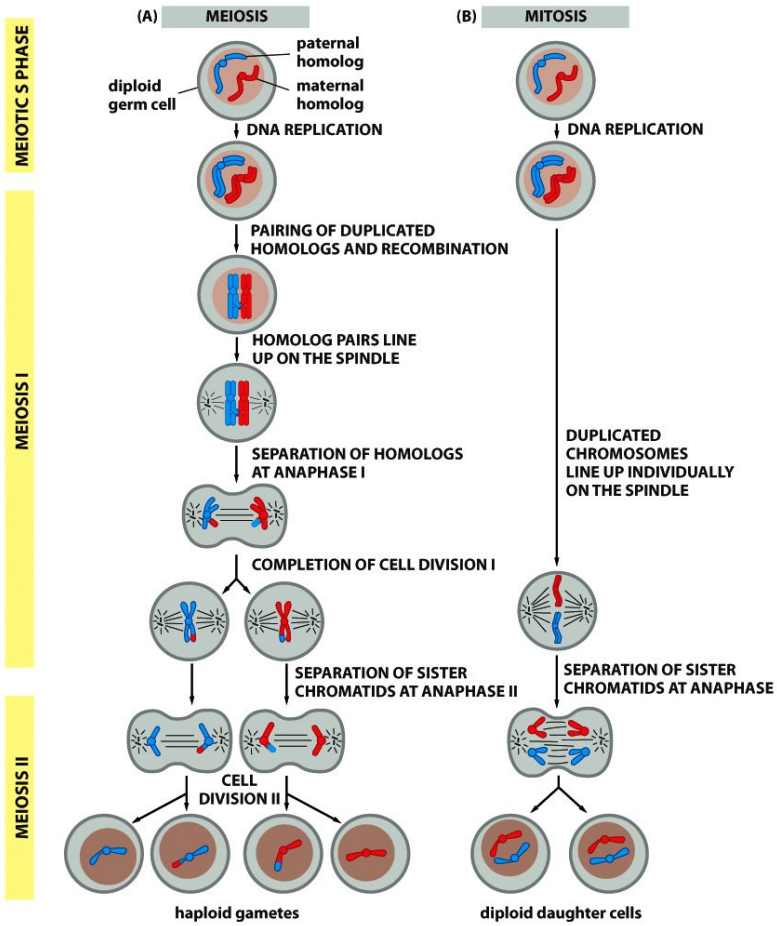


Figure 2.6 – Eukaryotic cell types and division. a) meiosis; b) mitosis. (Alberts et al., 2014)

The exchanged stretches of chromosome carry genetic information, thus some of the newly generated gametes will contain recombinant haplotypes (Clark & Pazdernik, 2012). Therefore, the same gene *locus* may contain different versions of the gene, depending on the nucleotide sequence of each piece of DNA. Finally, the alternative gene forms are denoted by the term alleles.

These forms may also dictate the outer appearance or functional role of a gene. Hence, some variants are dominant while others are recessive. The presence of the first type of mutation will most

often determine the trait expression. The only exception occurs when the phenotype is different from either one of those expressed by the dominant or the recessive alleles. Hence, this particular scenario denotes co-dominance of both alleles (Alberts et al., 2014).

These combinations of gene versions led to another form of classification. Therefore, subjects containing more than one allele type, for the same gene, are heterozygous while those containing identical copies of the same gene are homozygous (Alberts et al., 2014). The latter entails that an individual may carry solely the recessive or the dominant versions of the gene.

The crossover effect also depends on the physical distance between the genes. For instance, Figure 2.7c shows two chromosome homologs and the location of genes A, B, C, D, E, and F. Furthermore, the red chromosome contains only dominant genes (A – F) while the blue homolog contains the respective recessive forms (a – f). The short distance between genes A and B hinders the formation of a chiasma that would segregate the genotypes Abcdef and aBCDEF.

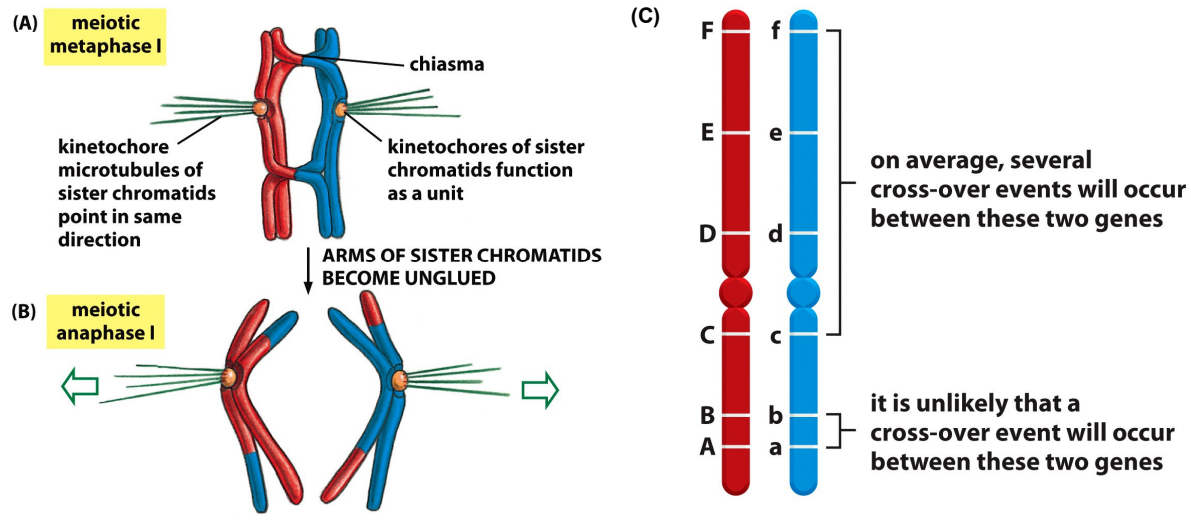


Figure 2.7 – Homolog pair (maternal-paternal) and gene locations. (Alberts et al., 2014)

The information contained in each allele serves as template for gene expression. Furthermore, the expression from a single allele is called Allele Specific Expression (ASE). The outcome of this expression may or may not be functional (Clark & Pazdernik, 2012). In fact, most functional products come from the dominant allele, since homozygous individuals, to the recessive allele, in most cases do not exhibit the phenotype or the functioning protein (Clark & Pazdernik, 2012).

The level of change, gain or loss of function depends on the extent of the nucleotide modification, since this variant might lead to deleterious alterations in the amino acid sequence of the target protein (Alberts et al., 2014; Branden & Tooze, 1999). Variants leading to non-synonymous changes may hinder the proper folding of the protein and ultimately affects its function (Branden & Tooze, 1999). However, synonymous mutations might still lead to functionally apt products. Finally, passing mutations might not even interfere in the function of the gene product.

These mutations are also classified according to the extent of the nucleotide change or base category. Thus, variation caused by the insertion or deletion of nucleotides are indels while a single change to the nucleotide identity is denoted by base substitutions (Alberts et al., 2014). Furthermore, a point mutation between purine (Adenine - Guanine) or pyrimidine (Cytosine – Thymine) bases is a transition while a change across base types consists of a transversion (Clark & Pazdernik, 2012).

These variants might affect the transcription rate of a functional gene and consequently the tissue protein levels. This inequality in protein production might lead to undesired consequences, for example: insufficient proteins to supply the tissue demand or enough molecules for effective gene regulation (Clark & Pazdernik, 2012). Therefore, the imbalance in allele specific expression is intimately linked with the description of the variants surrounding such genes.

There are many techniques that allow the genotyping of variants from an individual. For example, Short Nucleotide Polymorphism-array (SNP-array) (LaFramboise, 2009) and RNA-sequencing (Garber, Grabherr, Guttman, & Trapnell, 2011). The first gives the presence or absence of previously cataloged variants in genome of an individual while the second informs the researcher on the transcribed portions. Moreover, RNA-sequencing allows the detection of novel SNPs.

In essence, recombination explains the appearance of a diverse pool of chromosome configurations. This process shuffles the genetic information, hence it may also carry variants in coding and non-coding regions. It also determines the fitness of an individual, since it influences the expression of functioning gene products. In essence, these concepts lay the foundation for explaining trait inheritance.

2.1.2 Genetic Inheritance

Gregor Mendel postulated the laws of inheritance based on his study of pea phenotypes. However, that rationale adequately represents organisms that reproduce sexually (Alberts et al., 2014). The first rule consists of the law of segregation. It states that an organism segregates its genetic information into haplotypes to produce gametes and that two of these cells (one from each parent) randomly combine through fertilization (Laird & Lange, 2011). The second law is known as the law of independent assortment. It explains that during gamete formation the allele assignment to haplotypes occurs independent of the allele identity (Laird & Lange, 2011).

The haplotype formation through meiosis meets the requirements of Mendel's first law, since a diploid cell first duplicates its genetic content and then its daughter cells experience two subsequent cell divisions. This process produces cells with only one set of chromosomes, thus one copy of each gene (Alberts et al., 2014).

The gamete formation relies on pulling apart the chromatids of a chromosome. This process occurs at random, since there is no preferential direction for dividing the haplotype. Therefore, one chromosome has no influence on the separation of another. In other words, the genes of a chromosome do not impact on the segregation of the genes in another chromosome (Alberts et al., 2014). Finally, the

crossover events also ensure no selective dependence on the allele assignment to different haplotypes. Consequently, the combined result of these procedures theoretically abide to Mendel's second law (Laird & Lange, 2011).

However, the inheritance of variants, unlike genes, may defy the law of independent assortment. These mutations not always follow this rule because two variants that are spatially close to each other may be inherited together, thus the progeny displays a correlation in variant inheritance (Laird & Lange, 2011). Therefore, this phenomenon may explain the association between variants and a specific phenotype.

The inheritance of genes exhibiting allele specific expression could benefit from the analysis of the correlation between such variant pairs. Hence, these laws provide the framework for discussion of variant inheritance in terms of linkage disequilibrium.

2.1.3 Linkage Disequilibrium

Linkage Disequilibrium (LD) consists of the deviation from the rule of independent assortment proposed by Mendel. Therefore, pair-wise variants that seldom segregate during recombination are in high linkage disequilibrium (Mueller, 2004). In other words, they display high correlation. This behavior is best described by analyzing the hereditary nature of a variant after many generations.

In Figure 2.8, the common ancestor contains a marker (cross) in one of its chromosomes and this variant is bequeathed to the progeny. Furthermore, the authors (Laird & Lange, 2011) show mixture in the region before the marker, starting at the third generation. In fact, three out of seven of these individuals exhibit this type of mixture. Therefore, the recombination rate expresses the level of association between the two variants. If, the region is distant enough from the cross, then the recombination is an independent process and the expected recombination rate equals 0.5. Nevertheless, the opposite is also true, namely, variants located in a region that rarely experiences recombination will express it at a rate less than 0.5.

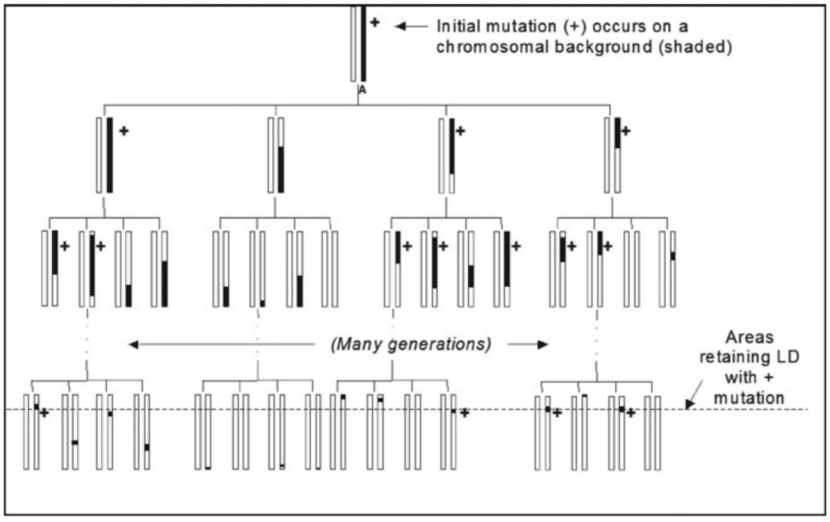


Figure 2.8 – Variants across many generations. (Laird & Lange, 2011)

This dichotomy exemplifies the null- and alternative-hypothesis in a test for linkage disequilibrium. Nonetheless, LD is best described by a coefficient that measures the deviation from independence in the inheritance of two variants (D). Moreover, the individual allele frequencies in a population influences this measurement, consequently the Pearson correlation (R or R^2) provides a stable measure to compare different variant pairs (Mueller, 2004).

In summary, variants in high LD display strong association ties due to chance or because they confer specific phenotypic traits. The pair-wise calculations of these coefficients are dependent on the allele frequencies of a population and small studies could borrow statistical power from data offered by online databases, such as: 1000G (Delaneau, Howie, Cox, Zagury, & Marchini, 2013), ENSEMBL (Yates et al., 2016) and SNAP (Johnson et al., 2008).

2.2 Functional Enrichment Analysis

The complex interactions in cell processes of eukaryotic cells employ many genes to accomplish biological processes and molecular functions. Therefore, functional enrichment analysis seeks to evaluate the statistical significance of annotation terms in describing the affected pathway (Wadi, Meyer, Weiser, Stein, & Reimand, 2016). The most common annotation sources consist of data bases describing the relationship between genes in the context of ontology terms (The Gene Ontology Consortium, 2000). There are different forms to evaluate functional enrichment (Huang, Sherman, & Lempicki, 2009). In fact, this section briefly describes three foundation methods: Singular Enrichment Analysis, Gene Set Enrichment Analysis, and Modular Enrichment Analysis.

Single Enrichment Analysis relies on proposing a list of genes and evaluating the membership of these genes to the individual annotation terms to arrive on a test statistic representing the term. This procedure leads to redundant annotation terms, since many biological processes are accomplished by groups of genes. In addition, the researcher must establish the significance threshold (Huang et al., 2009).

Gene Set Enrichment Analysis compares the annotation between two lists of genes and generates the threshold statistic based solely on the expression profiles of the data. The first consists of gene IDs and their corresponding expression level from up- to -down regulated entries while the second represents the genes belonging to a specific set, such as: an ontology term membership. This information provides the basis for calculation of the enrichment score, which represents the test statistic. Finally, the evaluation of the significance threshold of enrichment scores relies on random permutation of the subject labels (Subramanian et al., 2005).

These methods rely on calculating membership to a list or term without pooling information about shared annotations. Therefore, the Modular Enrichment Analysis tries to evaluate the membership statistic in terms of either sets of genes with a common term or terms with closely related genes (Huang et al., 2009). Moreover, this technique only requires one set of genes and an annotation source. Lastly,

the sharing of information across genes and terms reduces the redundancy found in the previous approaches.

There are many tools that apply and even expand these methods. However, the software is just as important as the annotation sources. Consequently, the investigation of functional enrichment should combine reliable tools with up-to-date databases. In fact, (Wadi et al., 2016) shows that the breadth of gene ontology terms has doubled between the years of 2009 and 2016. The authors also tracked the assimilation of these changes by the tools that rely on these sources. Consequently, the results of many software do not reflect these changes while others, such as ToppGene (Chen, Bardes, Aronow, & Jegga, 2009) and g:Profiler (Reimand, Kull, Peterson, Hansen, & Vilo, 2007), show considerable effort in presenting reliable analysis.

The primary goal of this analysis is to provide terms/sets that reflect enrichment among a pool of genes. Therefore, this technique could provide the biological processes involved in the list of genes generated by the algorithm proposed in this study.

3 MATERIALS AND METHODS

This study covers different data sources and a single model for shortlisting variants with potential regulatory role. In fact, Figure 3.1 describes the basic outline of this work. This section describes the assumptions and algorithms implemented in preprocessing the data, selecting variants and evaluating their function.

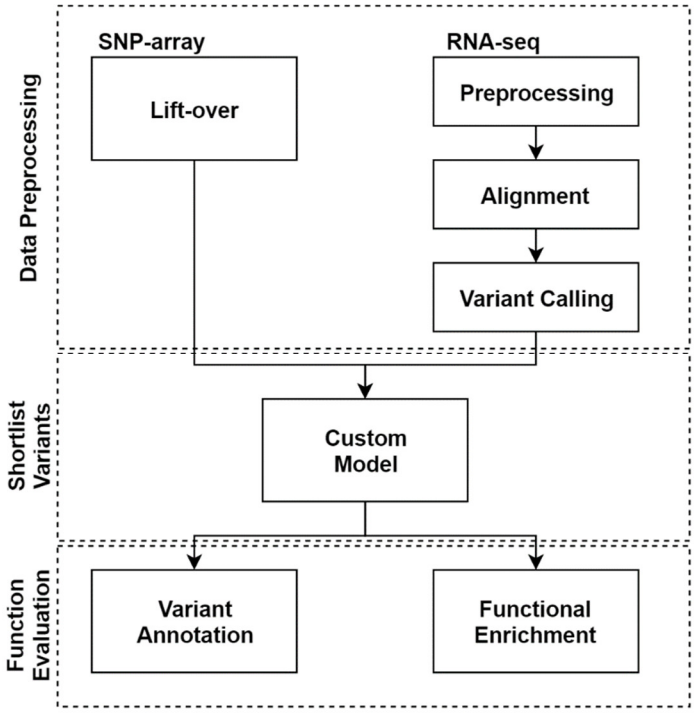


Figure 3.1 – Core sections illustrating the main tasks involved in this study.

3.1 Data Preprocessing

This analysis explores the information on eight European individuals under two different conditions. In other words, blood samples were collected from an even number of male and female individuals at their basal state, henceforth denoted by the term naïve, while subsequent treatment of the samples with lipopolysaccharide led to the stimulation of an immune response, thus this state is called stimulated or simply LPS.

The DNA content of the samples served as basis for genotyping through SNP-array while the RNA material provided the means for RNA-sequencing. In addition, (Edsgärd et al., 2016) describe the experimental procedure involved in performing these tasks. The authors investigated the same samples but with the aim of answering a different research question. Furthermore, the current work benefits from the quality control performed by the authors on the array data, thus it only utilizes variants that were deemed of high standard. Nonetheless, this study processed the sequencing reads by a different methodology.

3.1.1 SNP-array

The genotyping array provides information on loci of known variation. In fact, this study conducts analysis on the Illumina Omni BeadChip 2.5M micro-array. Furthermore, the original data is available at ArrayExpress with accession number E-MTAB-1450 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1450/).

(Edsgärd et al., 2016) processed this information and shortlisted high quality variants. However, that study was conducted in the context of the hg19 genome build while the current study is on GRCh38.p5. Therefore, the SNP-array had to be lifted from the hg19 reference to the current working version. Consequently, variants heterozygous in at least one sample had their coordinates updated by means of the CrossMap software (Zhao et al., 2014). This program takes as input a chain file as well as the Fasta file corresponding to the target genome build. The first was retrieved from the UCSC website at the following address: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz, while the second consists of the primary assembly of the genome sequence provided by GENCODE (GRCh38, ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_24/GRCh38.primary_assembly.genome.fa.gz).

The output of the software consists of the lifted coordinates with no information on the rsIDs of the selected variants. Therefore, these loci served as basis to retrieve the consensus reference nucleotide from the reference genome build via a custom Python script. The next step consisted of variant annotation with SnpEff (Cingolani et al., 2012). In other words, the reference allele was updated but the alternative allele remained the one reported by the array. Consequently, the retained variants are those with evidence supporting the occurrence of the mutation - dbSNP v. 144 (Sherry et al., 2001). The selected variation database is housed at the National Center for Biotechnology Information - NCBI (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/VCF/GATK/All_20160407.vcf.gz).

3.1.2 RNA-sequencing

The sequencing data provides a measure of gene expression in terms of read counts. However, these samples present Illumina HiSeq 2000 paired-end reads with an average length of 100 bp at varying sequencing depths. Therefore, the analysis requires read quality control prior to alignment to the reference genome. This section presents the preprocessing measures as well as the parameters for read mapping and variant calling.

3.1.2.1 RNA-sequencing Preprocessing Steps

This stage consists of a two-step approach to filter the data and perform quality control. The unprocessed files are stored on the Sequence Read Archive (SRA-NCBI) with accession number SRA062051 (http://www.ncbi.nlm.nih.gov/sra/?term=SRA062051). The initial procedure consisted of a test to determine the level of sample contamination by ribosomal RNA. Consequently, SortMeRNA v. 2.1 (Kopylova, Noé, & Touzet, 2012) filtered the reads according to the following databases:

Table 3.1– Ribosomal RNA Databases

Database	Ribosomal Unit	Filter (minimum overlap, %)
Silva (Bacteria) (Quast et al., 2013)	16S	90
	23S	98
Silva (Archaea) (Quast et al., 2013)	16S	95
	23S	98
Silva (Eukarya) (Quast et al., 2013)	18S	95
	28S	98
RFAM (Griffiths-Jones, Bateman, Marshall, Khanna, & Eddy, 2003)	5S and 5.8S	98

The second step involves the removal of sequencing adapters and the trimming of reads failing quality control. These tasks were achieved with the Trimmomatic (v. 0.36) software. In essence, the algorithm removed the TruSeq adapter from the Illumina data. Furthermore, it also evaluated the nucleotide sequence quality at 5 bp windows and trimmed reads that presented bases with a quality below 20 at any given window. A sample command with all of the quality parameters along with their definition is available on **Appendix A**.

Finally, each biological sample was loaded across multiple flow cells of the sequencing machine. Therefore, data preprocessing and mapping occurred on a lane dependent manner to detect and treat biases related to sequencing artifacts.

3.1.2.2 Read mapping

The choice of aligner was based on performance as well as mapping accuracy. In fact, the evaluation performed by (Engström et al., 2013) greatly contributed to this decision, since it narrowed the number of options to two software, namely: GSNAP and STAR. Furthermore, GATK also suggest the use of STAR in their best practices guidelines for RNA-sequencing (Auwera et al., 2014; DePristo et al., 2011). Therefore, STAR v.2.5.1b was selected as the aligner of this study.

The software aligned the reads to the reference genome. However, prior to alignment it requires the creation of an index based on the genome build and on information about splice junctions. The latter corresponds to the gene annotation version 24 from GENCODE (ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_24/gencode.v24.annotation.gtf.gz). **Appendix A** presents further information on the command arguments for generating the reference index files.

The aligner processed the operation in the two pass mode (--twopassMode Basic) and converted the quality scale from Phred-33 to Phred-64 (--outQSconversionAdd -31). In addition, it introduced the XS attribute for unstranded data into the output file. It is also worth mentioning that the reference Fasta file included scaffold regions in order to provide an enriched region with highly occurring ribosomal RNA. Lastly, **Appendix A** holds the command summarizing this information.

3.1.2.3 Variant Calling

The differential expression from the alleles of heterozygous variants helps on the identification of genes that could potentially display allele specific expression. Therefore, the aligned reads were introduced to a variant calling procedure strongly influenced by the suggestions of the GATK Best Practices (Auwera et al., 2014; DePristo et al., 2011).

Figure 3.2 presents a flowchart containing the steps of this pipeline, from lane-dependent aligned reads until individual read counts for the selected variants. Furthermore, this figure depicts the software and main parameters corresponding to each task.

The initial step consists of adding read group information to enhance the performance of the haplotype caller. Therefore, Picard Tools (The BROAD Institute, 2016) includes this data as well as merges the lane-dependent files and discards duplicate reads. In addition, the merger retains the information on individual lanes, thus allowing proper quality score recalibration.

The Split'N'Trim routine modifies the reads by reassigning the maximum sequencing quality of good alignments as well as trimming reads spanning splice junctions. The latter prevents the occurrence of false positives by not allowing overhangs while the former allows the base recalibrator to better estimate the mapping error rate (parameter: reassign mapping quality from 255 to 60, RMQF 255 RMQT 60).

The next step in this process is indel realignment. Firstly, the algorithm creates a list of target *loci* that exhibit potential sequencing artifacts and could benefit from a novel alignment. Finally, these *loci* undergo *de novo* alignment with the aid of information from a supporting indel database. This file

consisted of the database compiled by the Broad Institute containing the data from (Mills et al., 2006) and the 1000 Genomes Project (Auton et al., 2015). The source is available at ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/hg38bundle/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz.

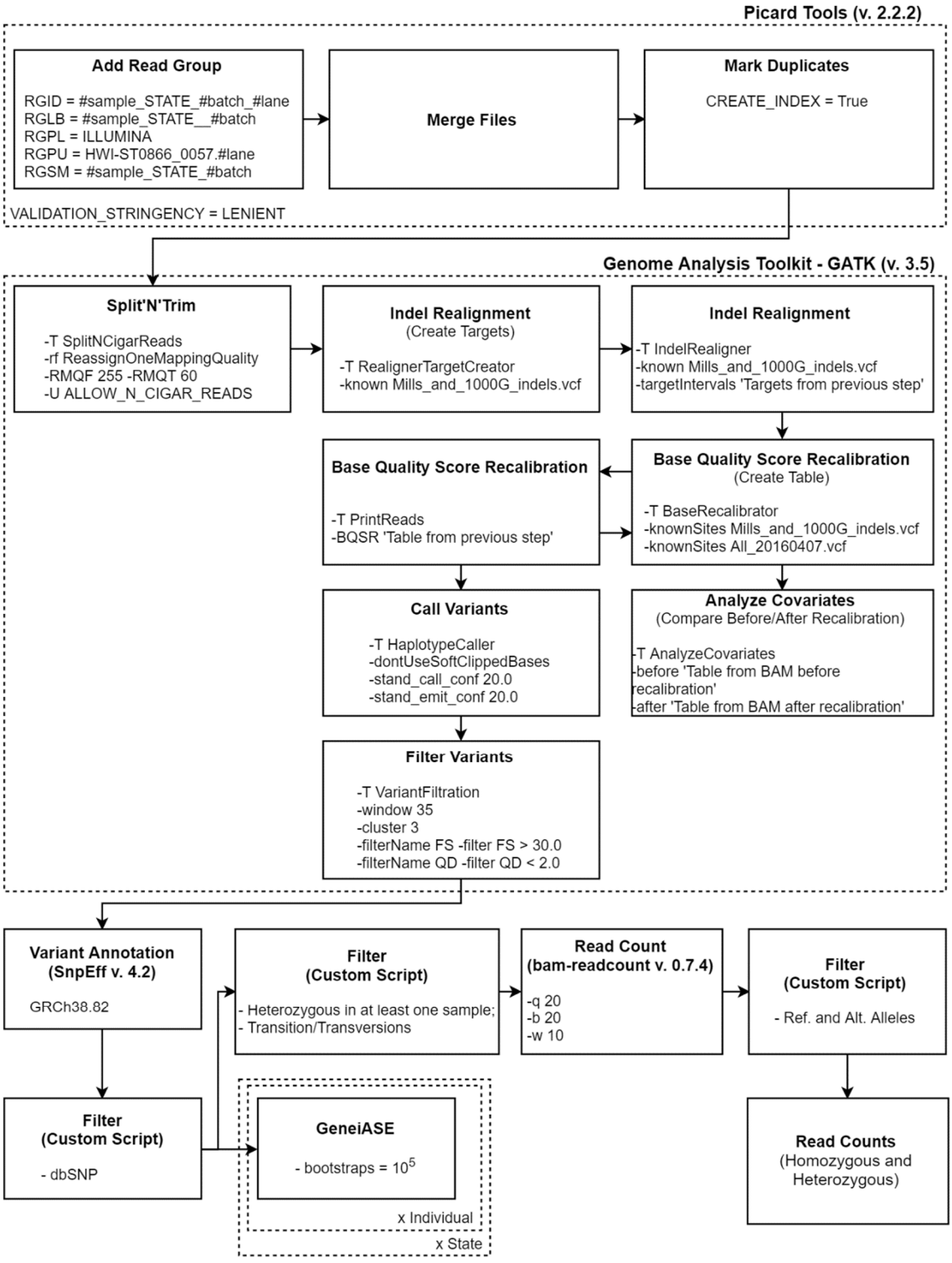


Figure 3.2 – Variant Calling Pipeline

Another measure to improve the reliability of the data is to perform base recalibration. This process uses a machine learning algorithm to correct the imbalance between reported and empirical read quality. The software models the empirical error in the nucleotide assignment of bases outside of known variation *loci* (dbSNP v.144; and Mills et al., 2009), then applies a correction to the reported values, in order for them to better represent the empirical mismatches. In addition, the tool Analyze Covariates determines the efficiency of this step and generates a summary describing the read quality before and after the procedure.

The main tool of the variant calling pipeline is the haplotype caller. This algorithm is designed to detect the variation with respect to the reference (GRCh38.p5) while disregarding or minimizing the influence posed by sequencing and mapping artifacts. In fact, GATK suggests the use of a Phred score of 20 for both confidence parameters: calling and emitting variants.

The final step on detecting variation involves the filtering of the emitted entries. This task increases the confidence of the called variants by analyzing the Fisher Strand value ($FS > 30$) and the Quality by Depth parameter ($QD < 2.0$). The Fisher Strand statistic denotes the Fisher exact test applied to evaluate strand bias (forward/reverse), either on the reference or on the alternative allele, while the QD consists of the read quality normalized by the sequencing depth. The latter prevents enlarged quality scores that are caused by an increased coverage. In addition, this algorithm also filters clusters containing 3 or more variants in a 35 base window.

The filtered variants could represent known or novel SNPs. Therefore, the next procedure corresponds to annotate the variants with the aid of SnpEff (Reference: GRCh38.82) while the subsequent step filters the data for known variants (rsIDs, dbSNP v. 144). This pool of mutations allows the estimation of the potential of a gene to exhibit ASE. To this end, the GeneiASE software (Edsgärd et al., 2016) performed this evaluation with an overdispersion coefficient of 0.49, a permutation number of 10^5 (bootstraps) and variants with at least 10 reads across alleles (indels and base substitutions).

So far, the data has been processed on a sample and lane dependent level, hence the following step concatenates the information from the individuals that share the same condition (naïve/LPS). A custom Python script selects base substitution variants, then shortlists mutations that are heterozygous in at least one sample. It also restricts the choice of heterozygous variants to those containing a minimum of 10 reads across alleles.

Nonetheless, this study requires read counts not only of heterozygous individuals but also of homozygous subjects. For instance, if a heterozygous variant encodes a base substitution from A to G, then the model asks for the read counts from individuals that are homozygous to the A-allele (reference), homozygous to the G-allele (alternative) as well as the number of reads from the heterozygous subjects containing a G as the alternative allele. Moreover, this study performed variant calling at an individual level, as suggested by the GATK best practices. As a result, the data was not pooled across samples and allele counts were reported only for the variants of the individual under analysis.

Consequently, the software Bam-Readcount - version 0.7.4, (Larson & Abbott, 2016) - provided the total number of reads mapping to the shortlisted *loci.* The main parameters consisted of minimum base quality (b) and minimum mapping quality (q). Therefore, the selected value coincides with the criteria stipulated for the Haplotype caller (20). In addition, no restriction was placed on the number of reads mapping to this configuration.

The resulting read count includes all nucleotides meeting the consistency criteria, hence another Python script stores the number of reads corresponding to the reference and alternative allele of the selected rsIDs. This operation suffices to present the custom model with enough information on the distribution of sequences of each sample. Therefore, the next section discusses the model assumptions and the mathematical framework.

3.2 Shortlisting Variants

The proposed algorithm takes read counts from RNA-sequencing as well as genotype information from the SNP-array and performs a statistical test to shortlist exon variants that are indicative of ASE-genes. These variants serve as anchors for pairwise linkage disequilibrium analysis. Therefore, the following stage consists of matching exon SNPs with intron and upstream variants for functional enrichment investigation. Figure 3.3 depicts the general work flow of the model.

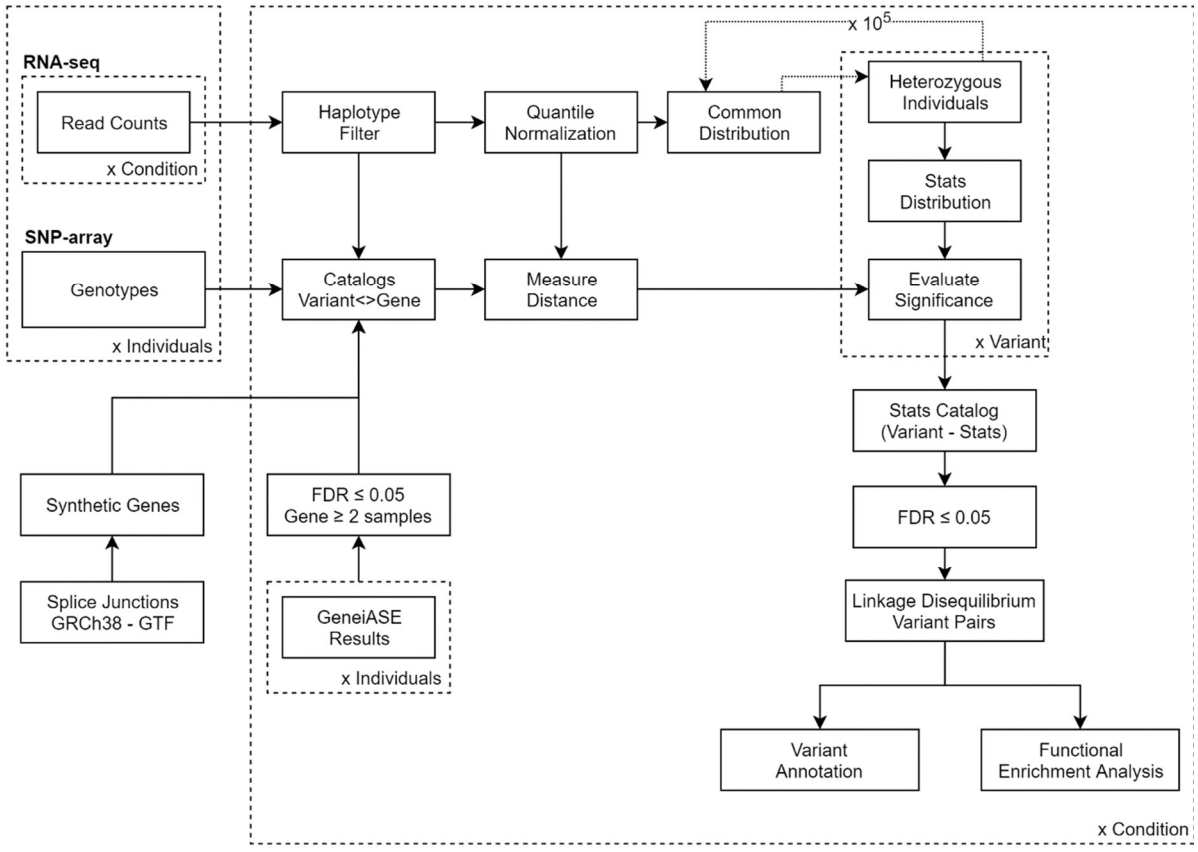


Figure 3.3 – Work flow of the model to shortlist RNA-sequencing variants and generate SNP-pairs of ASE genes for functional analysis.

This study uses the concept of synthetic gene boundaries to classify the variants according to exon, intron or upstream regions. This construct consolidates the gene isoforms in a simplified splice junction set. Therefore, the procedure merges exons across isoforms, if they present an overlap. The resulting boundaries assures that intron/upstream variants do not overlap with any exonic region from the different isoforms. The original gene annotations correspond to the splice junctions from GENCODE v.24. This file was processed in a custom Python script. In addition, the construct interprets the beginning of the gene as the leftmost coordinate of its first exon. Figure 3.4 exemplifies the merging process.

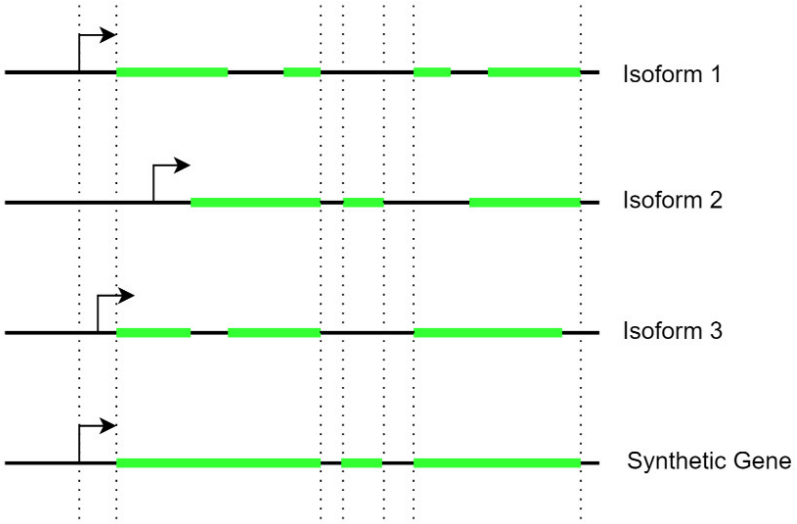


Figure 3.4 – Illustration of the merging procedure to generate a synthetic gene boundary.

The work restricts the analysis to genes that show potential allele specific expression. Therefore, the results from the GeneiASE software (Edsgärd et al., 2016) play an important role in selecting the genes eligible for this analysis. Therefore, successful candidates must present a False Discovery Rate (FDR) less than or equal to ≤ 0.05. In addition, the need of information on all three haplotypes restricts this list to genes present in at least 3 samples.

The RNA-sequencing variants are also subject to restrictions, such as: a haplotype-like filter. This measure ensures that the selected variants have at least one heterozygous individual as well as one individual for each of the homozygous haplotypes (reference and alternative). The surviving variants along with those from the genotyping array are cataloged according to gene and region (exon, intron or upstream).

The inter-individual analysis required a common read count measure. Thus, this approach borrows the quantile normalization method from (Bolstad, Irizarry, Astrand, & Speed, 2003). In addition, individuals lacking read counts for a specific variant have their values replaced by the sample

median, only for this calculation, since the missing data symbol was reassigned to these variants after normalization.

The data on homozygous occurrences was halved based on the assumption of equal chromosome contribution towards total gene expression. The underlying motivation for this approach relies on the notion of chromosome dosage (Clark & Pazdernik, 2012), since genes presenting ASE might provide insufficient gene products on its biological setting. Consequently, halving the expression of homozygous variants presents a more leveled comparison. It also greatly improves the shortlisting process by introducing the possibility of classifying the haplotype contribution in different haplotype-plots (Figure 3.5; Figure 4.10 - Figure 4.12).

The normalization procedure renders a common read count distribution for all samples of the same state (naïve/LPS). In addition, this transformation generated a continuous variable well suited for an inverted Weibull distribution.

3.2.1 Custom Model

The model evaluates the shape of the variant plot as well as determines the significance of the distance measure supporting this shape, in order to shortlist exon SNPs. The graph heavily relies on the read counts for each haplotype, since it depicts the mean expression of the homozygous samples as well as the individual counts of the alternative allele for each heterozygous occurrence. In essence, the introduction of the haplotype-plot allows the comparison of biological data against a null model without introducing bias towards extreme values based solely on read counts.

The initial step on this evaluation consists of determining the plot shape of the biological data. The accepted forms are: rising, falling, step-up, step-down, flat, concave, convex, nconcave, and nconvex. Figure 3.5 illustrates these formats while sample data is available on Figure 4.10 - Figure 4.12 of the Results section. The step-up and step-down form factors are special cases of the rising and falling shapes while the nconcave and nconvex represent non-symmetrical concave or convex plots, respectively.

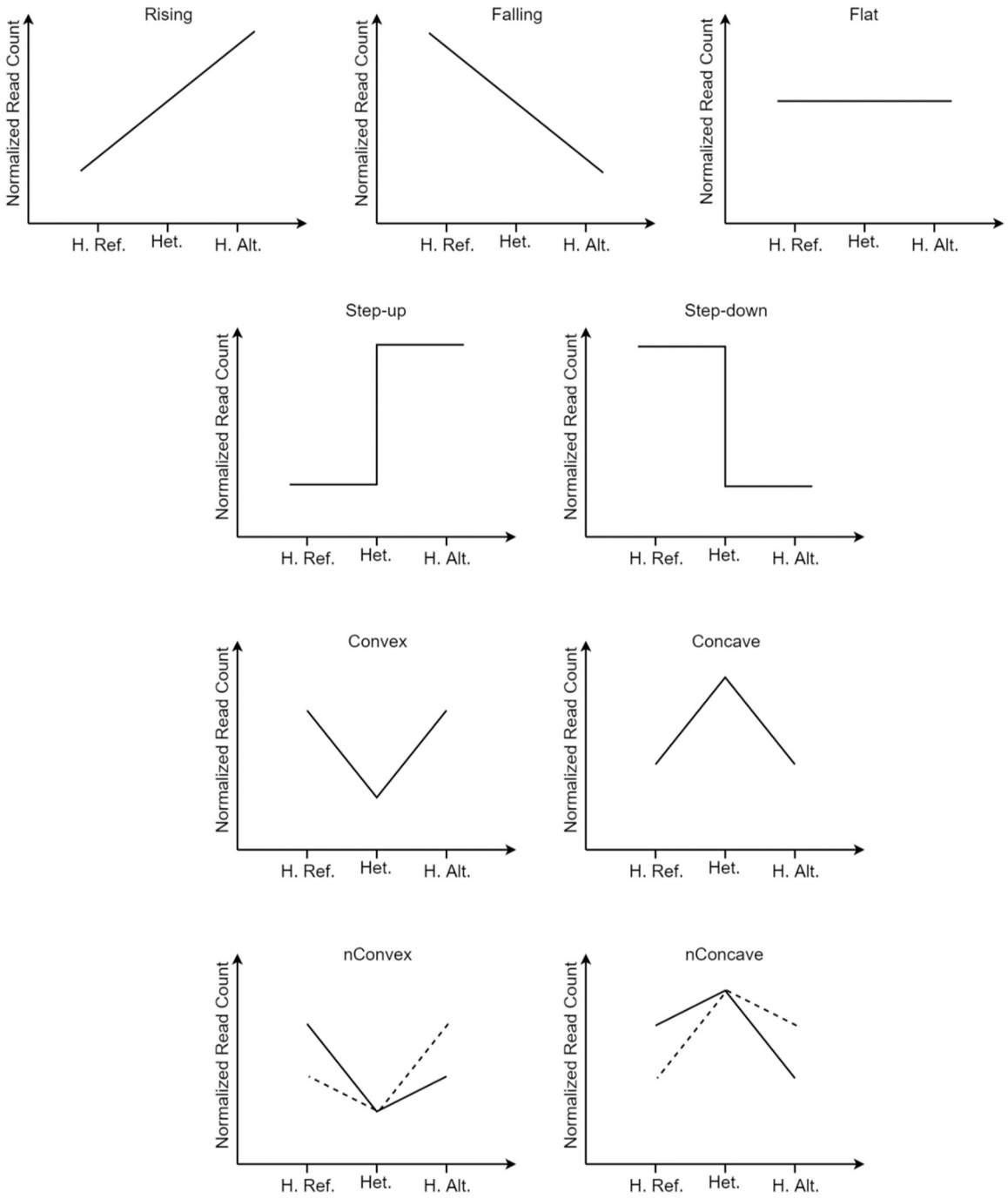


Figure 3.5 – Allowable configurations for the haplotype-plots.

Notes:
- H. Ref. – Average read count from the individuals Homozygous to the Reference allele
- Het. – Heterozygous Individual (alternative allele)
- H. Alt. – Average read count from the individuals Homozygous to the Alternative allele
- The dotted line shows an alternative plot shape for the given scenario.

The flat category denotes variants showing little difference from the behavior of the homozygous individuals. Furthermore, this definition classifies a variant as “flat” if the expression level of all individuals (hom/het) is within 2 normalized read counts. Such instances are not further processed, since they are not considered to be representative of allelic imbalance.

This threshold also guides the decision on assigning the convex/concave shape, since these plots are regarded as symmetrical with respect to the y axis (Normalized Read Count). Nonetheless, the restriction in expression level is confined only to the homozygous samples.

Another factor influencing the determination of plot shape is the presence of outliers. This study handles these situations according to haplotype: homozygous or heterozygous. Initially, the read counts are selected based on the variant under analysis, then the detection of outliers within the homozygous samples follows method A while those in the heterozygous configuration follows method B. The first approach evaluates the distance between the individual reads to the haplotype median (Equation 1) and discards those exhibiting more than a two-fold deviation from the median distance of the centered data (Equation 2). In addition, the information on the selected samples collapses to a single value represented by the mean. Therefore, after outlier detection and removal, the model calculates the average number of sequences from each homozygous haplotype.

$$x_{i,j}^{Hom} = abs\left(rc_{i,j}^{Hom} \quad median\left(rc_{i}^{Hom}\right)\right) \qquad \textit{Equation 1}$$

$$s_{i,j}^{Hom} = \frac{x_{i,j}^{Hom}}{median\left(x_{i}^{Hom}\right)} \qquad \textit{Equation 2}$$

rc – Read Counts;
i – Variant under analysis;
j – Samples sharing the same haplotype configuration for the current variant;
Hom – Homozygous individuals (alternative and reference are evaluated separately); and,
x – Distance from the reads to the centered data around the homozygous haplotype median.

The second method evaluates the plot shape for each heterozygous individual and exon variant, then it counts the votes towards the different plot shapes. Hence, the selection criterion becomes membership to the pattern with the highest number of votes under the analyzed variant. In other words, method B excludes those individuals not contributing to the selected pattern from downstream analysis of the given exon-variant. In this sense, the graph best represents those samples carrying a more pronounced allelic dependent expression.

The test statistic depends on the distance between the heterozygous samples and the mean read count from one of the homozygous haplotypes. Thus, it can be measured with respect to the reference (d^{+}) or the alternative (d^{*}) homozygous configuration, as depicted in Figure 3.6 and Equation 3. As a result, the model evaluates the distance to both haplotypes and chooses the anchor based on the majority rule.

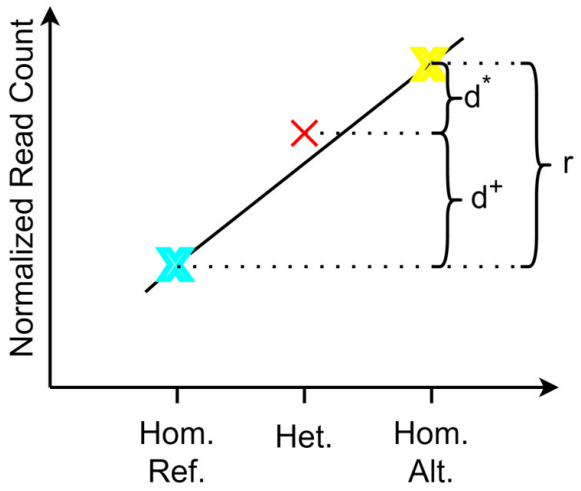


Figure 3.6 – Haplotype plot illustrating the distances between the average read counts of the homozygous individuals and one heterozygous sample.

Equation 3 illustrates the distance from the heterozygous samples to the homozygous subjects with respect to the possible haplotype configurations. It takes one of two forms depending on the anchor.

$$d^{+}_{i,j} = rc_{i,j} \quad avg\left(rc_i^{Hom.Ref}\right)$$
$$d_{i,j} = rc_{i,j} \quad avg\left(rc_i^{Hom.Alt}\right) \qquad \textit{Equation 3}$$

rc – Read Counts;
i – Variant under analysis;
j – Heterozygous individual;
Hom. Ref – Samples homozygous to the reference allele;
Hom. Alt – Samples homozygous to the alternative allele; and,
d – Distance from the heterozygous sample to the average homozygous haplotype.

The summarization of this measure depends on a modified average, similar to the Stouffer's method. Firstly, the data is divided by the sample standard deviation, in order to better control the variance. If the distances are very similar, then the model establishes a lower bound of 10^{-6} to prevent a ratio overflow. Secondly, the corrected distances ($\tilde{d}_{i,j}$) are averaged by the square root of the number of heterozygous individuals (N). These steps summarize the information on the heterozygous samples of the given variant and are depicted on Equation 4 and Equation 5.

$$\tilde{d}_{i,j} = \frac{d_{i,j}}{\sqrt{\frac{1}{N \quad 1}\sum_{j=1}^{N}\left(d_{i,j} \quad \bar{d}_i\right)^2}} \qquad \textit{Equation 4}$$

$$V_i = \frac{\sum_{j=1}^{N}\tilde{d}_{i,j}}{\sqrt{N}} \qquad \textit{Equation 5}$$

$d_{i,j}$ – Distance from the heterozygous sample to the average homozygous haplotype (reference or alternative, according to the majority vote);
$\tilde{d}_{i,j}$ – Corrected distances;
i – Variant under analysis;
j – Heterozygous individual; and,
N – Total number of heterozygous samples containing variant *i*.

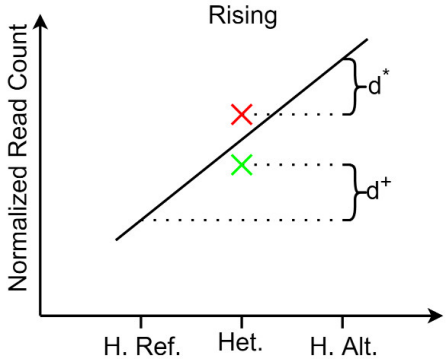
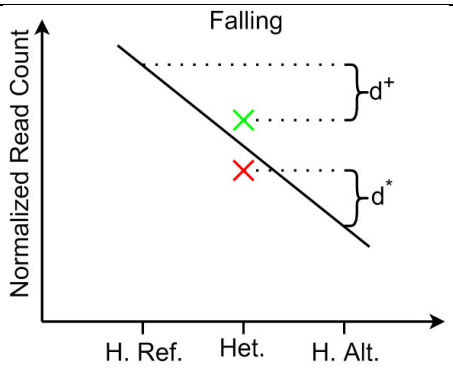
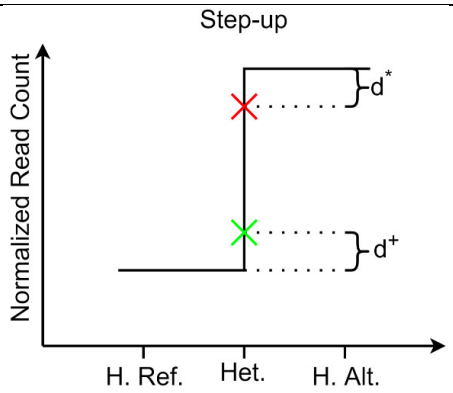
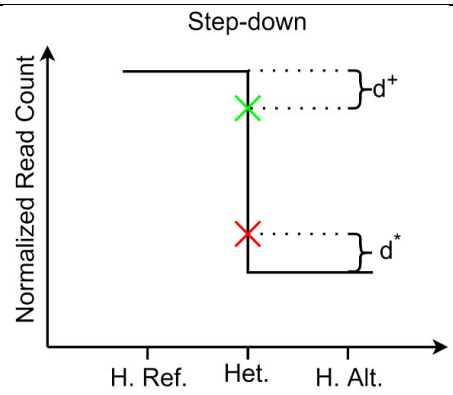
This statistic (V_i) compared against a distribution generated under a null model enables the evaluation of the significance of the variant to ASE. In fact, the common distribution obtained from the quantile normalization provides the means for sampling read counts. Overall, the model randomly samples 10^5 data points from the common distribution for each heterozygous individual at every exon variant. This mutation has its own plot shape, as obtained from the biological data. Therefore, extreme values are counted according to the patterns outlined in Table 3.2.

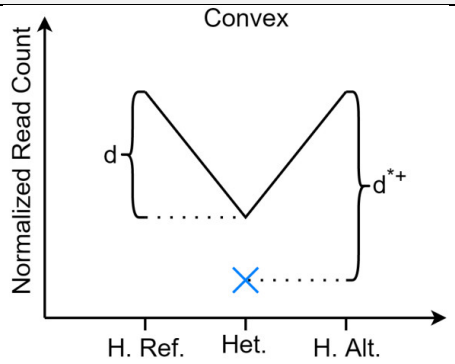
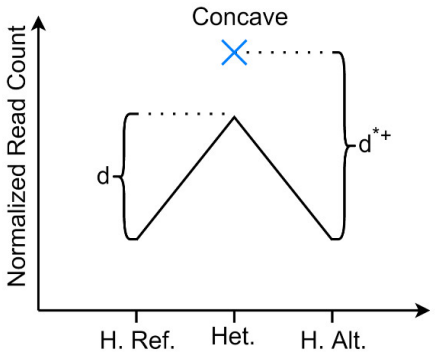
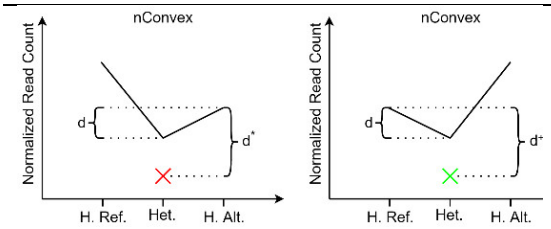
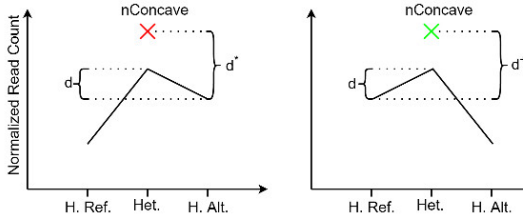
The haplotype anchor determines the sign on the distance measure as well as the regions containing values as extreme or more extreme than the biological data. For example, Table 3.2 depicts the case of a single heterozygous sample with different plot patterns as well as anchors (red cross/*, green cross/+). In addition, d represents the distance found in the biological data while d^* and d^+ represent the convention of Equation 3. Therefore, if the plot shape shows a rising trend with anchor on the average read count of the samples homozygous to the reference allele (+), then all distances with respect to the anchor are positive. Moreover, the null model will generate a similar or more extreme result only if d^+ is less than the value obtained from the biological sample but greater than zero.

In essence, there are two groups of patterns: the linear-shapes and the v-shapes. The first harbors the rising, falling, step-up, and step down patterns while the second contains the concave, convex, nconcave and nconvex. Therefore, extreme values for the linear-shapes are taken as those falling as close or closer to the anchor, since they show an intermediate expression when compared to individual alleles from the homozygous configurations. On the contrary, the v-shapes exhibit a heterozygous contribution that is higher (concave/nconcave) or lower (convex/nconvex) than the averages of any homozygous haplotype. Consequently, the extreme values of these patterns are either as distant or more distant than the value found in the biological data.

Overall, a statistic is only considered as extreme or more extreme than the biological data if it corresponds to one of the scenarios outlined on Table 3.2. Consequently, each variant carries a p-value and its calculation follows the guidelines of (Phipson & Smyth, 2010). Finally, the last step on shortlisting exon variants constitutes the calculation of the False Discovery Rate (Storey & Tibshirani, 2003).

Table 3.2 – Possible scenarios and extreme value regions.

Plot Shape	Alternative Anchor $d_{i,j} = rc_{i,j} \quad avg(rc_i^{Hom.Alt})$		Reference Anchor $d_{i,j}^{+} = rc_{i,j} \quad avg(rc_i^{Hom.Ref})$	
	Sign	**Extreme Region**	**Sign**	**Extreme Region**
	-	$0 \geq d^{*} \geq d$	+	$0 \leq d^{+} \leq d$
	+	$0 \leq d^{*} \leq d$	-	$0 \geq d^{+} \geq d$
	$d \approx 0$	$0 \geq d^{*} \geq d$	$d \approx 0$	$0 \leq d^{+} \leq d$
	$d \approx 0$	$0 \leq d^{*} \leq d$	$d \approx 0$	$0 \geq d^{+} \geq d$

Plot Shape	Alternative Anchor $d_{i,j} = rc_{i,j} \quad avg(rc_i^{Hom.Alt})$		Reference Anchor $d_{i,j}^{+} = rc_{i,j} \quad avg(rc_i^{Hom.Ref})$	
	Sign	**Extreme Region**	**Sign**	**Extreme Region**
	-	$d^{*} \leq d$	-	$d^{+} \leq d$
	+	$d^{*} \geq d$	+	$d^{+} \geq d$
	-	$d^{*} \leq d$	-	$d^{+} \leq d$
	+	$d^{*} \geq d$	+	$d^{+} \geq d$

3.2.2 Linkage Disequilibrium

The linkage disequilibrium analysis depends on pairing the shortlisted exon variants with SNPs on intronic and upstream regions. Therefore, all mutations surviving an FDR filter of 0.05 were paired for further analysis.

The sample size of this study is insufficient for direct LD calculation. Thus, this work borrows the 1000 Genomes Project sample information on the European population (v. 5 – 20130502; CEU, TSI, FIN, GBR, and IBS). The files containing this information are available at: ftp://ftp-

trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/supporting/bcf_files/. In addition, the PLINK v. 1.9 software (Chang et al., 2015) greatly assisted in this analysis.

The raw data from the 1000G required preprocessing before LD calculation. Hence, the first step consisted of filtering the BCF files specifically for the samples of European individuals. The next procedure corresponded to marking duplicated variants, followed by their removal and creation of BED and FAM files. As a result, this pipeline generated one set of binary files per chromosome.

The final LD evaluation also benefited from the use of PLINK. In fact, the main parameters correspond to the calculation of the correlation for every input variant up to 100 SNPs away as well as no lower limit on the Pearson correlation. **Appendix A** presents a sample command for this procedure.

Finally, a custom Python script recovers the pair information and shortlists intronic and upstream variants for evaluation of regulatory role. In addition, the surviving transitions/transversions present a correlation higher than or equal to 0.8. Finally, the selected SNPs proceed to functional evaluation.

3.3 Functional Evaluation

3.3.1 Variant Annotation

The variant pairs presenting high correlation show strong signs of allele specific expression. Furthermore, this study concerns the evaluation of their potential regulatory role. Therefore, the Variant Effect Predictor (McLaren et al., 2010) provides reliable annotation for the variants on the intronic/upstream regions. In other words, this tool indicates if the variant overlaps known regulatory regions.

3.3.2 Functional Enrichment Analysis

The ToppGene Suite (Chen et al., 2009) contributed to evaluate functional enrichment of the affected genes, more specifically the tool “ToppFun: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis”. It performed the analysis on two categories: gene ontology terms and metabolic pathways.

The selected gene ontology categories consist of molecular function, biological process and cellular component. Furthermore, the pathway analysis included the following databases: BIOCYC (Caspi et al., 2014), KEGG (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016; Ogata et al., 1999), NCI - Pathway Interaction DB (Schaefer et al., 2009), Reactome (Croft et al., 2014), Wiki Pathways (Kutmon et al., 2015), GenMAPP - Gene Map Annotator and Pathway Profiler(Salomonis et al., 2007), MSigDB C2 BIOCARTA v5.1(Subramanian et al., 2005), Panther DB (Thomas et al., 2003), and SMPDB – Small Molecule Pathway Data Base (Frolkis et al., 2009; Jewison et al., 2014). Lastly, the tool compares the genes against a maximum annotation size of 2000 genes per term, a p-value cutoff equal to 0.05 and reports the results within an FDR of 0.05.

Finally, ambiguities in the enriched terms were solved by running the g:Profiler tool (Reimand et al., 2007) on functional enrichment mode and with the standard parameters.

4 RESULTS

4.1 Data Preprocessing

4.1.1 SNP-array

The original data consisted of 2.5M loci from the Illumina Omni BeadChip. However, this number considerably decreased with the filtering processed by (Edsgärd et al., 2016). Therefore, the number of eligible variants totals 1,019,730. Nonetheless, the lift-over procedure further reduced this number to 282,402, which corresponds to a 72.31% loss.

4.1.2 RNA-sequencing

4.1.2.1 Preprocessing

Table 4.1 summarizes the number of reads before and after the preprocessing steps.

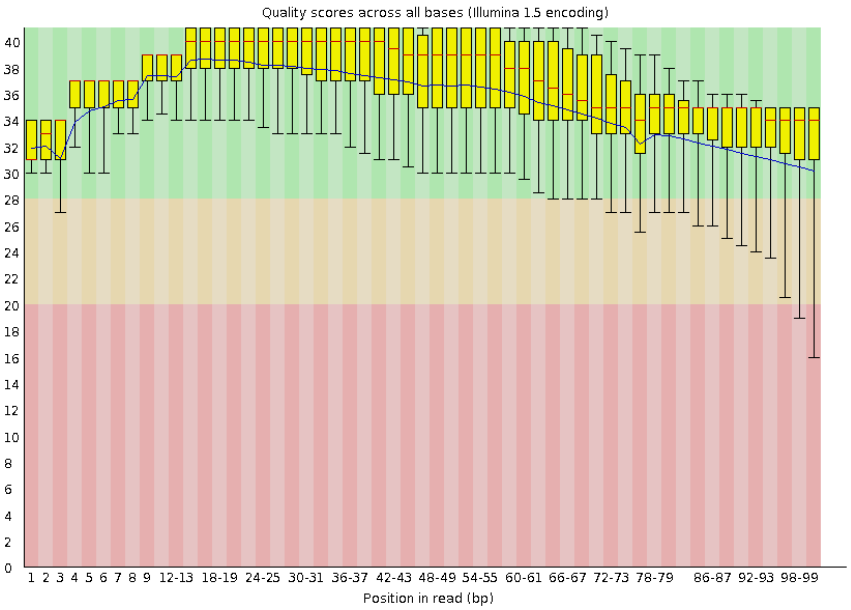
Table 4.1 – Number of reads before and after quality control, according to sample and condition.

Sample*	Sex	Age	Naïve			Stimulated		
			Before (x 10^6)	After (x 10^6)	Reduction (%)	Before (x 10^6)	After (x 10^6)	Reduction (%)
1	Female	27	52.499	35.050	33.24	83.210	56.504	32.09
2	Female	35	66.152	45.017	31.95	115.678	80.579	30.34
3	Male	39	86.855	63.221	27.21	90.043	74.280	17.51
4	Male	47	87.405	79.545	8.99	72.872	66.877	8.23
6	Female	30	51.310	47.089	8.23	93.107	85.073	8.63
7	Male	36	38.623	34.107	11.69	133.536	116.984	12.40
8	Male	30	37.028	33.244	10.22	195.610	175.129	10.47
9	Female	33	71.668	49.992	30.25	55.240	37.440	32.22

* There is no sample 5.

The read quality of the initial sequencing files shows very similar behavior across samples. Therefore, Figure 4.1 depicts the untreated state (Sample 1, Lane 1) while Figure 4.2 illustrates the treated scenario (Sample 2, Lane 2). These figures contain two parts: ‘a’ and ‘b’. The first illustrates the range of scores found in the pool of sequences of each sample and lane while the second depicts the empirical and theoretical GC content. Moreover, both reports are from the FastQC software (Andrews, 2010).

a)



b)

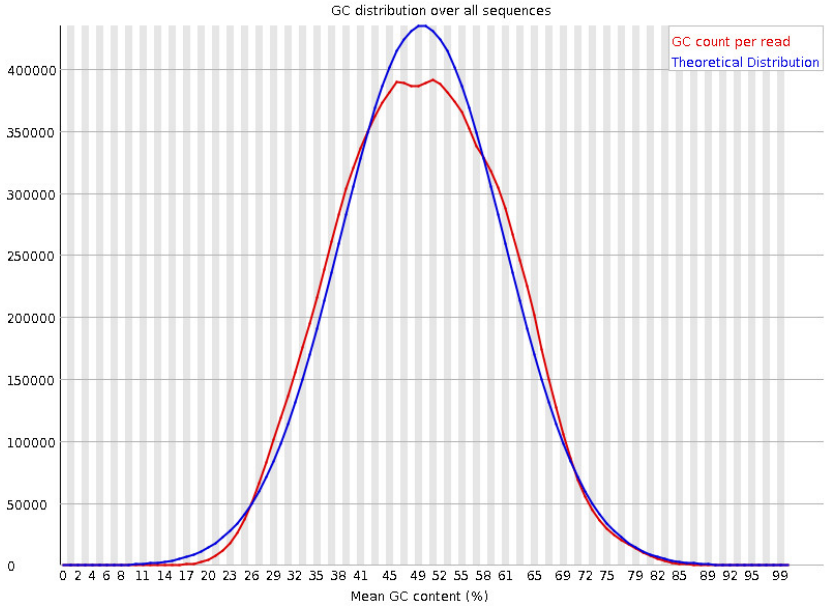
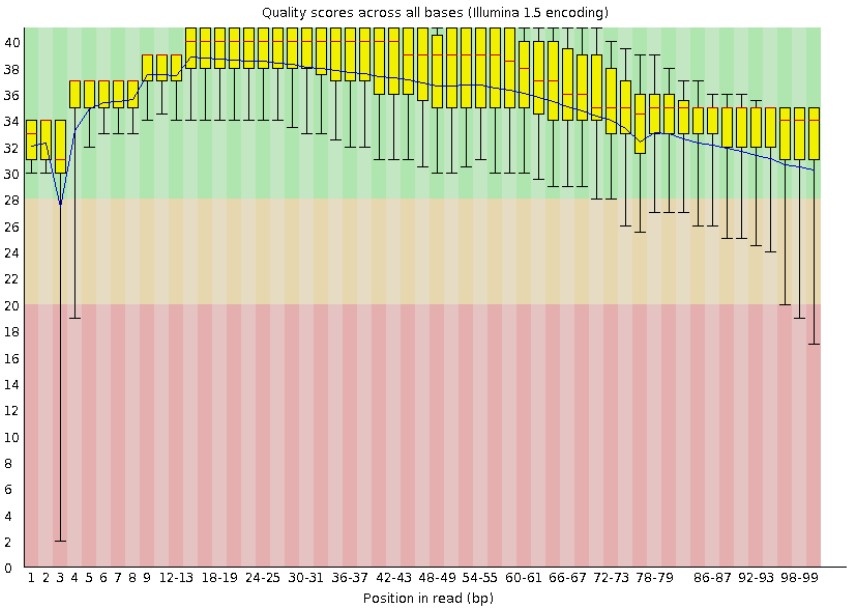


Figure 4.1 – Quality assessnent of the original data - Sample 1 (naïve state): (a) read quality as a function of read length and (b) GC content distribution.

a)



b)

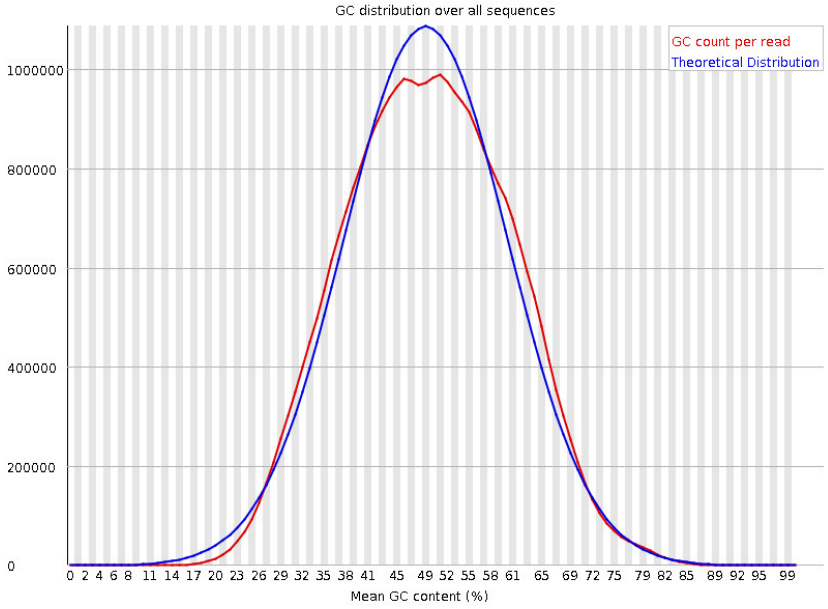
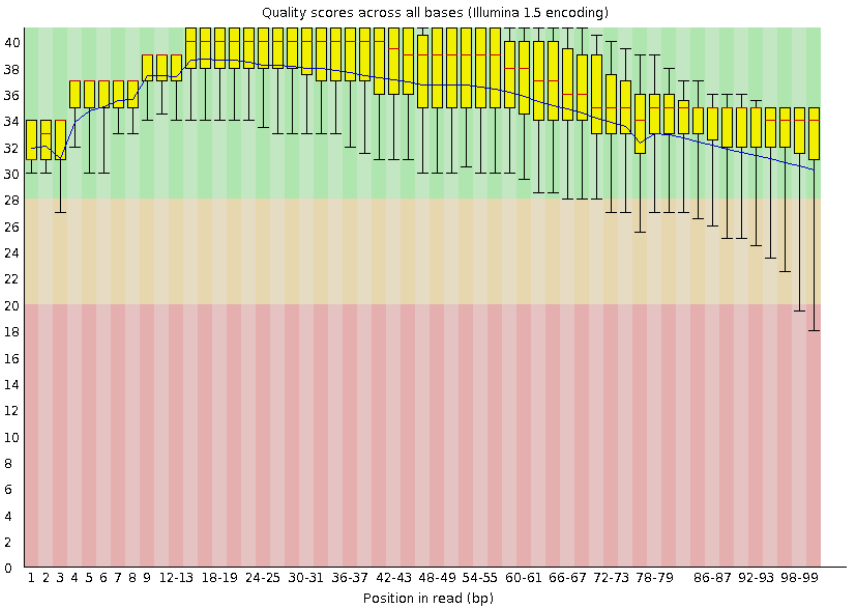


Figure 4.2 – Quality assessment of the original data - Sample 2 (LPS state): (a) read quality as a function of read length and (b) GC content distribution.

The effect of the removal of ribosomal RNA from the study samples also closely follows the trend of the raw files. Hence, sample 1 (Figure 4.3) represents the naïve state and sample 2 (Figure 4.4) the cells under an immune response.

a)



b)

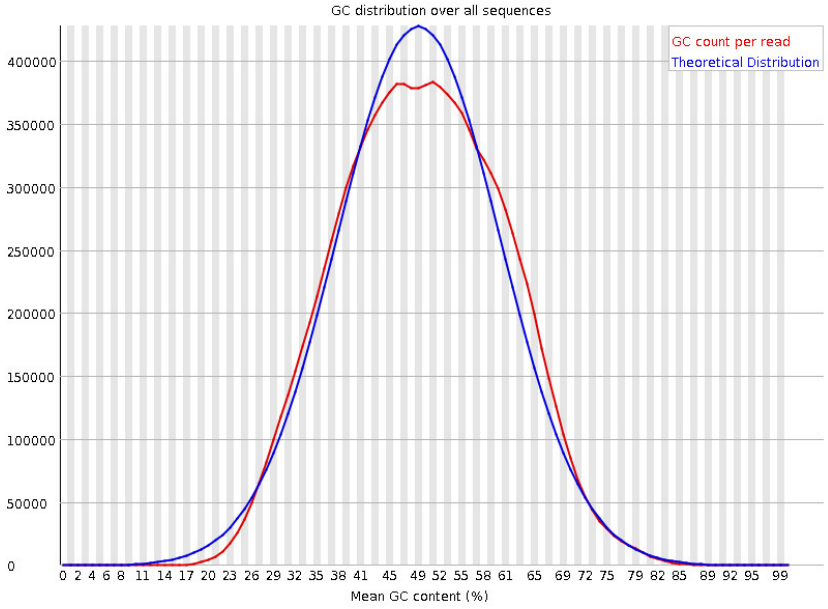
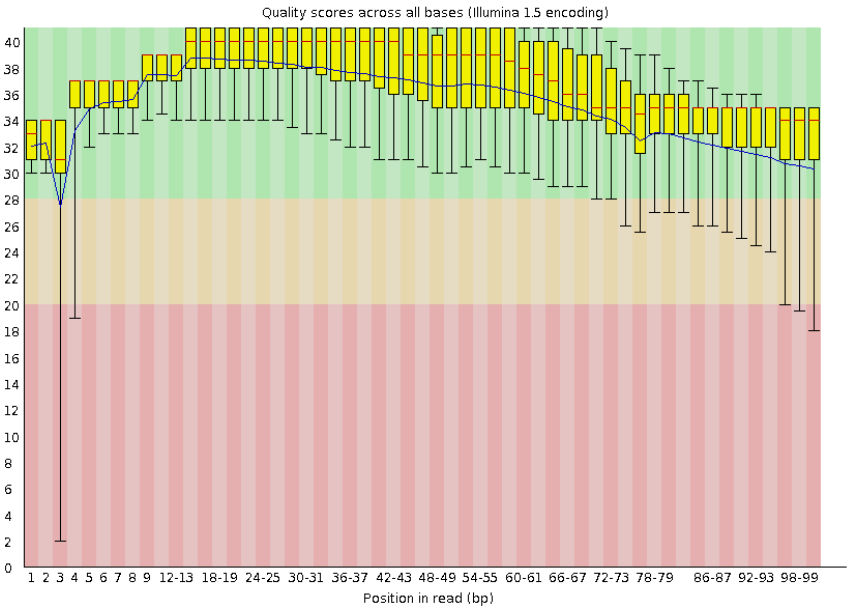


Figure 4.3 – Quality assessment of the data after ribosomal RNA removal - Sample 1 (naïve state): (a) read quality as a function of read length and (b) GC content distribution.

a)



b)

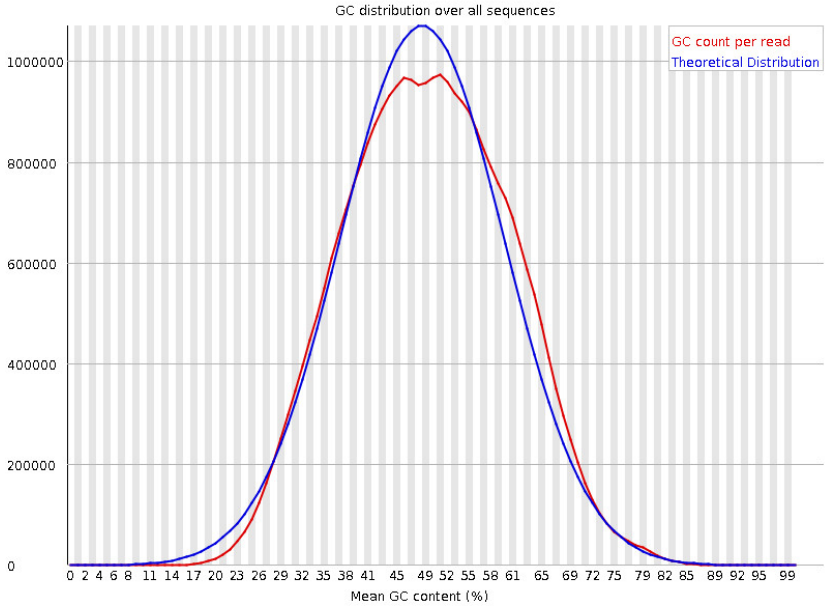
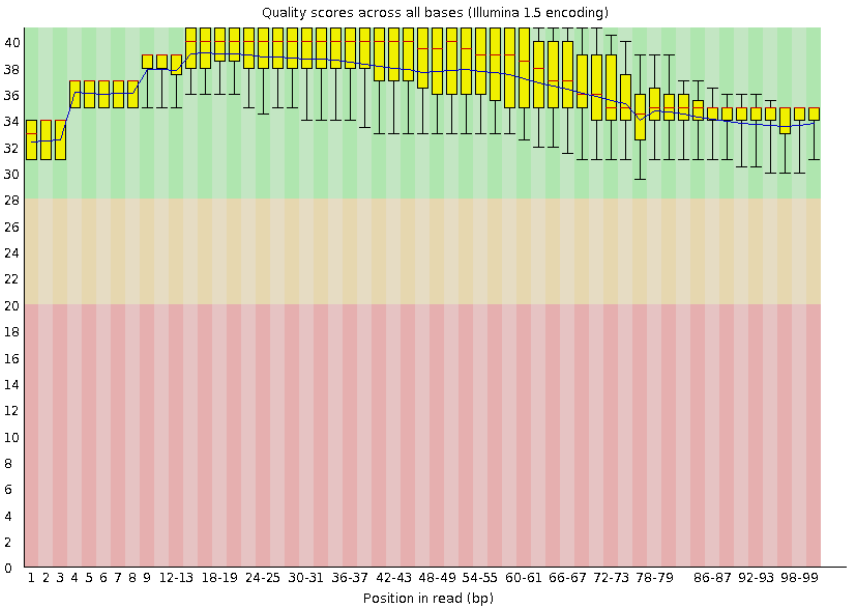


Figure 4.4 – Quality assessment of the data after ribosomal RNA removal - Sample 2 (LPS state): (a) read quality as a function of read length and (b) GC content distribution.

Trimming the reads leads to considerable change in these plots, as can be seen on Figure 4.5 (unstimulated) and Figure 4.6 (stimulated).

a)



b)

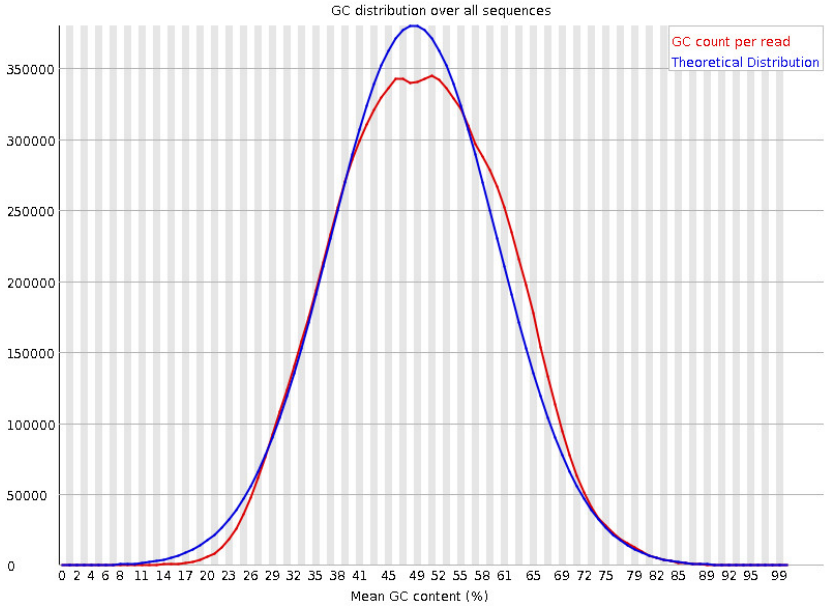
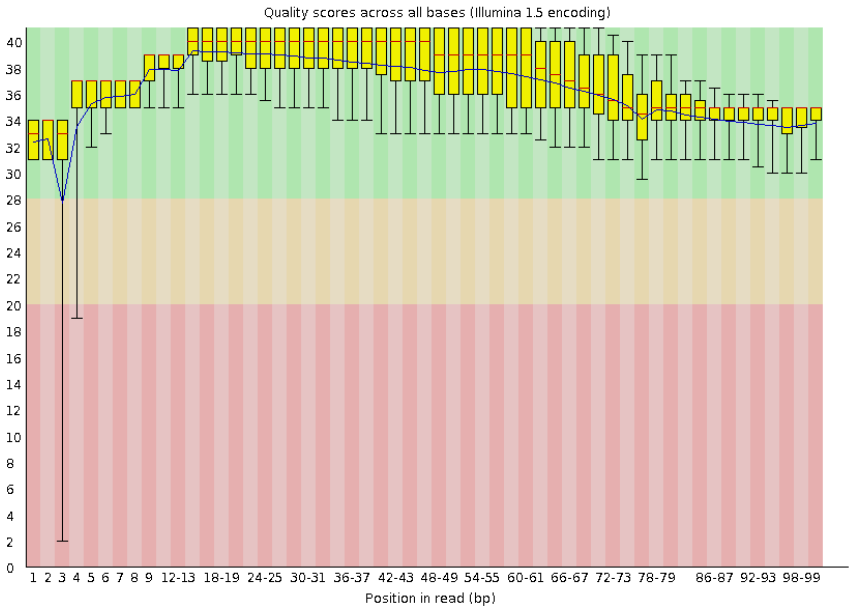


Figure 4.5 – Quality assessment of the data after trimming reads - Sample 1 (naïve state): (a) read quality as a function of read length and (b) GC content distribution.

a)



b)

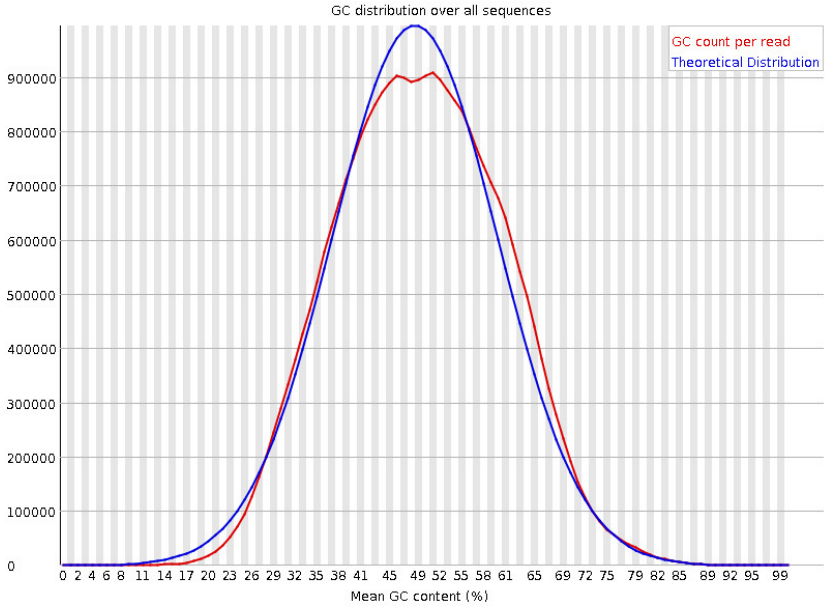
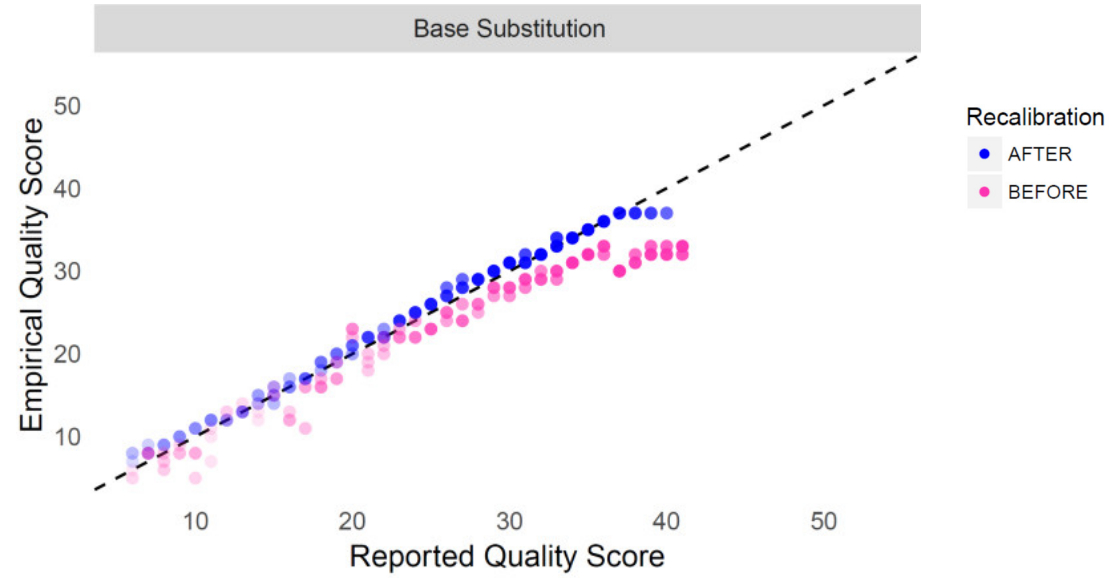


Figure 4.6 – Quality assessment of the data after trimming reads - Sample 2 (LPS state): (a) read quality as a function of read length and (b) GC content distribution.

4.1.2.2 Variant Calling

Figure 4.7 and Figure 4.8 depict the distributions of base quality score before and after base recalibration.

a)



b)

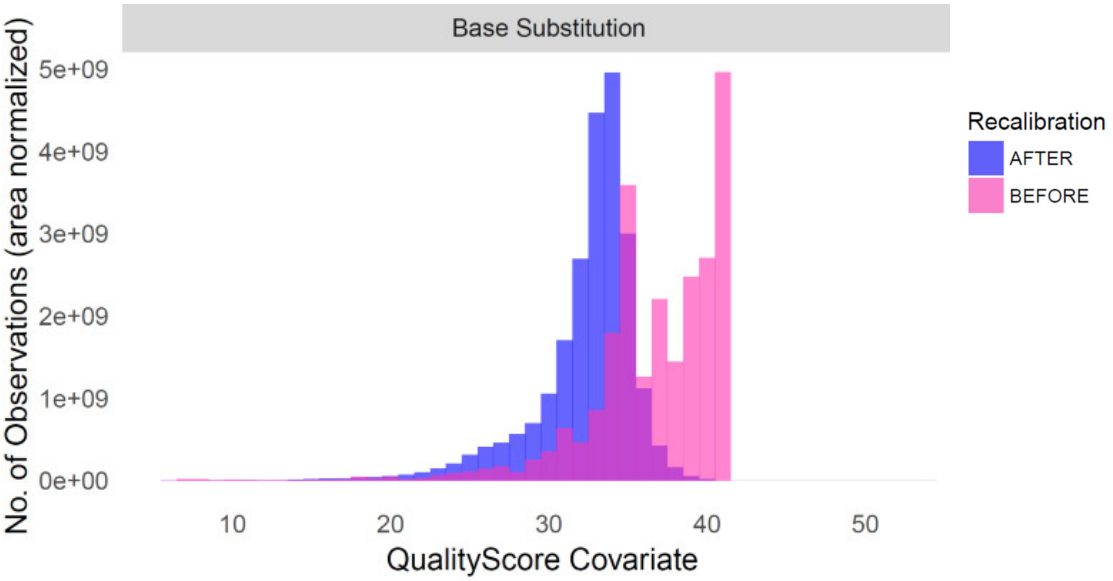


Figure 4.7 – Base Quality Score Recalibration report of sample 1 (naïve). a) Empirical vs. reported quality Score; b) Nucleotide distribution by quality score.

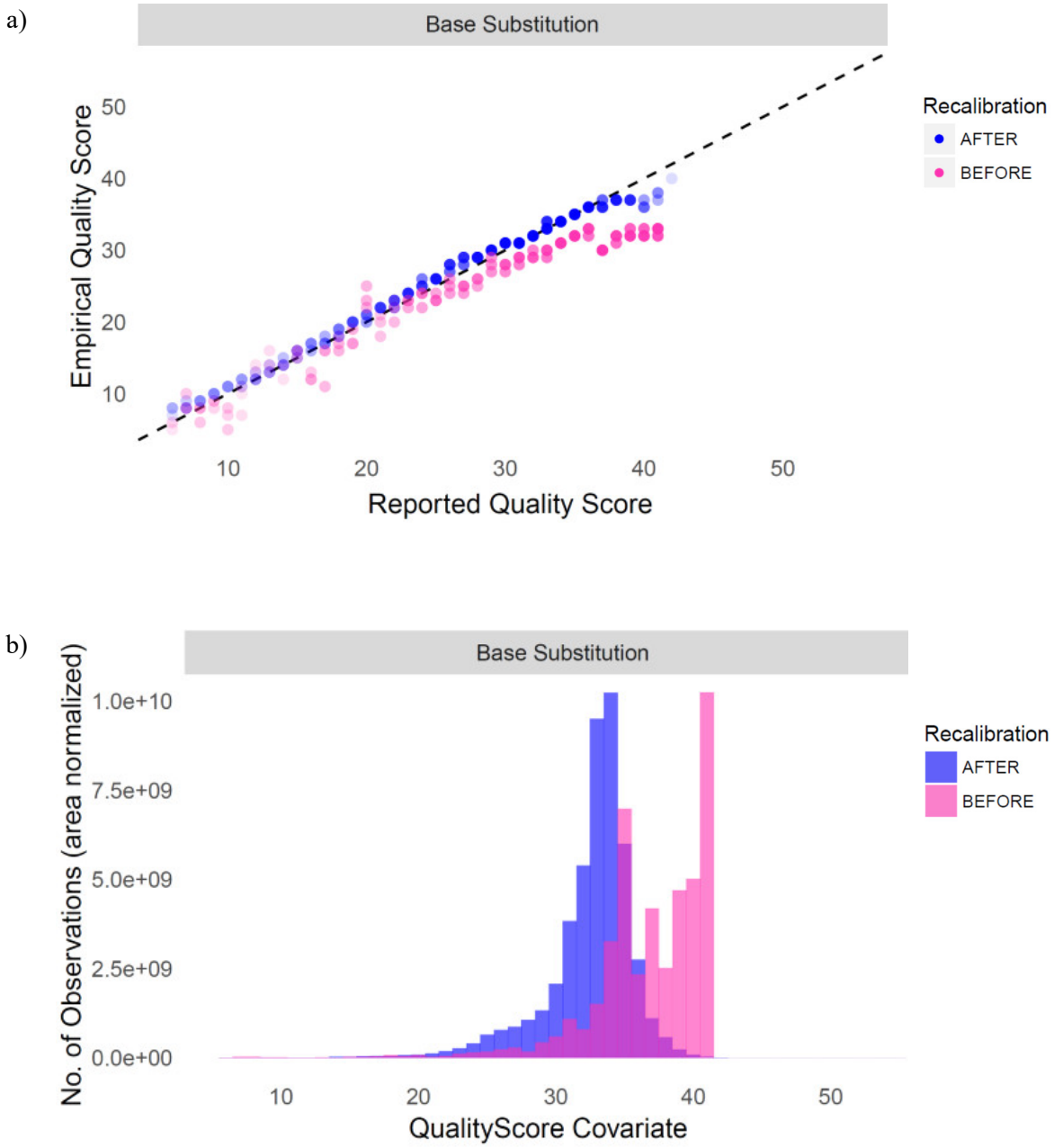


Figure 4.8 – Base Quality Score Recalibration report of sample 2 (LPS). a) Empirical vs. reported quality Score; b) Nucleotide distribution by quality score.

Table 4.2 presents the total number of variants obtained from the variant calling pipeline for the untreated samples while Table 4.3 depicts the data of the stimulated state. Furthermore, they present data on three filters: quality, type and haplotype. The quality filter consists of variant calls with high confidence while the type filter restricts the analysis to base substitutions. Lastly, the haplotype filter presents the number of variants containing at least one individual in each haplotype configuration. In other words, at least one subject of the following haplotypes: heterozygous, homozygous to the reference, and homozygous to the alternative allele. Moreover, the last filter requires the heterozygous individuals to exhibit a read count across alleles of at least 10.

Table 4.2 – Number of variant calls from the naïve samples before and after filtering for high call quality, variant type and haplotype configuration.

Sample	Variant Calls	High Quality Calls	Base Substitutions	Haplotype Filter
1	224,752	194,055	150,744	45,745
2	248,299	208,400	154,699	45,975
3	299,454	245,714	178,738	46,283
4	313,729	265,910	205,224	46,515
6	244,286	209,292	161,459	46,168
7	166,724	147,252	115,605	44,917
8	200,024	174,617	136,038	45,698
9	262,232	220,652	160,060	46,149
			Overall	**46,911**
			Unique*	**10,367**

*unique entries are those exclusive to this state.

Table 4.3 – Number of variant calls from the LPS samples before and after filtering for high call quality, variant type and haplotype configuration.

Sample	Variant Calls	High Quality Calls	Base Substitutions	Haplotype Filter
1	239,222	201,528	153,801	63,835
2	345,939	282,801	207,201	65,786
3	271,369	225,794	168,861	64,784
4	261,293	220,424	169,192	65,025
6	339,048	284,241	218,023	66,355
7	251,381	208,171	158,167	63,146
8	513,838	412,607	299,798	67,143
9	213,634	186,051	149,146	64,122
			Overall	**67,638**
			Unique*	**31,094**

*unique entries are those exclusive to this state.

4.1.2.3 Selection Criteria for ASE Genes

The following two tables depict the number of genes surviving the following selection criteria: FDR and minimum number of individuals simultaneously expressing the gene.

Table 4.4 - Number of eligible genes in the naïve scenario at varying FDR and sample membership levels.

FDR	# Individuals			
	3	4	6	8
0.01	321	238	113	18
0.05	535 (unique* = 144)	405	195	39
0.10	723	550	256	53

*unique entries are those exclusive to this state.

Table 4.5 - Number of eligible genes in the stimulated scenario at varying FDR and sample membership levels.

FDR	# Individuals			
	3	4	6	8
0.01	541	437	265	144
0.05	884 (unique* = 493)	674	419	238
0.10	1191	920	570	314

*unique entries are those exclusive to this state.

The synthetic gene boundaries consist of 60,517 gene definitions. These splice junctions served as guidelines to classify the variants. Therefore, Table 4.6 concatenates this information with the genes from Table 4.4 and Table 4.5 (3 individuals and FDR ≤ 0.05) as well as the variants from Table 4.2 and Table 4.3.

Table 4.6 – Number of variants detected by RNA-seq as a function of state and category (exon, intron, and upstream).

Category	Naive State	Stimulated State
Exon	1,823	2,849
Intron	3,783	8,563
Upstream*	1,188	2,104
Total	6,750	13,354

*These *loci* may overlap with intron variants, hence the total is not the simple sum across categories.

4.1.2.4 Quantile Normalization

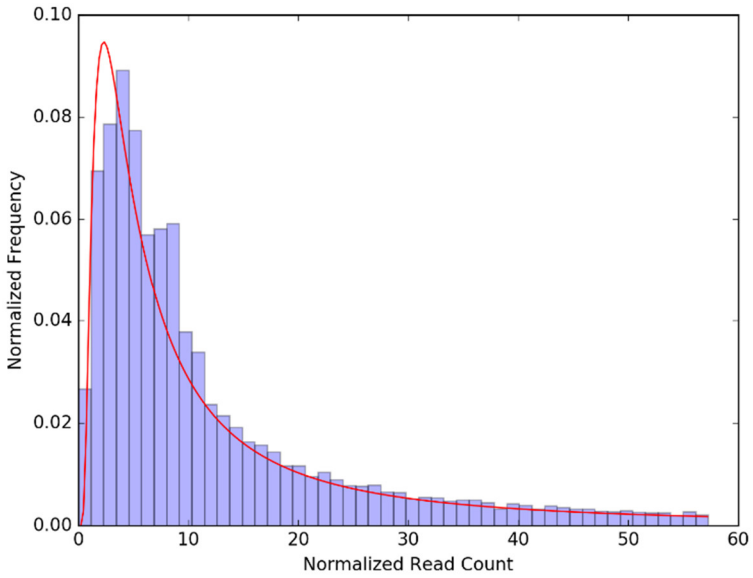
Table 4.7 presents the parameters obtained from fitting the normalized data to an inverted Weibull distribution.

Table 4.7 – Parameters of the fit of an Inverted Weibull distribution to the quantile normalized data, according to state.

State	Shape	Location	Scale
Naïve	0.8733	0	5.5465
Location	0.8210	0	4.8482

Figure 4.9 depicts the normalized read count data as well as the fitted distributions for both scenarios.

a)



b)

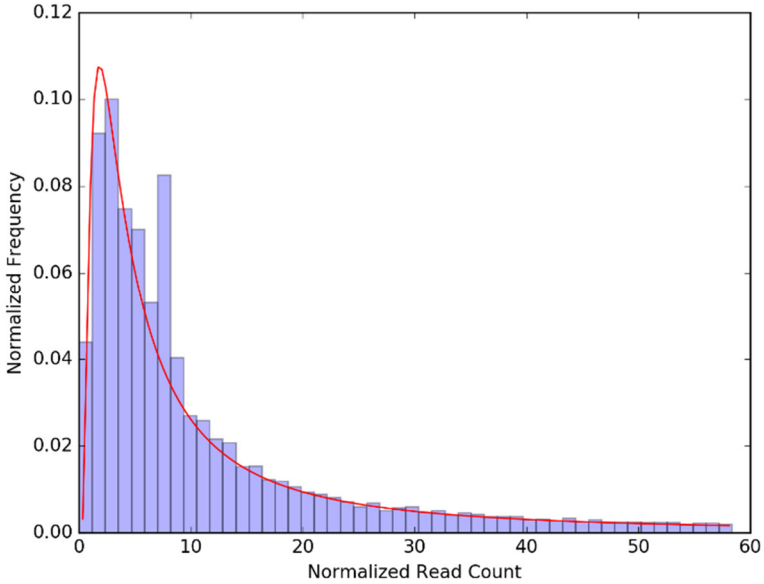


Figure 4.9 - Normalized read counts fitted to an inverted Weibull distribution a) unstimulated; and, b) stimulated states.

4.2 Shortlist Variants

4.2.1 Custom Model

Table 4.8 summarizes the patterns of the haplotype plots for all exon variants containing nearby intronic/upstream variants.

Table 4.8 – Number of haplotype plots classified by pattern and state without restriction on FDR.

Pattern	Naive State	Stimulated State
nConcave	437	741
nConvex	273	541
Falling	240	397
Rising	206	307
Concave	156	193
Convex	43	63
Inconclusive	291	417
Flat	74	79
Total	**1720**	**2738**

The aim of this calculation is to provide a list of exon-intron/upstream variant pairs. However, the number of exon variants without nearby intronic/upstream mutations corresponds to 103 for the naïve samples and 111 for the stimulated scenario. In addition, the number of base substitutions in exon regions with an FDR ≤ 0.05 is equal to 234 (naïve) and 740 (LPS). The latter leads to the following haplotype plot counts.

Table 4.9 – Number of haplotype plots classified by pattern and state (FDR ≤ 0.05).

Pattern	Naive State	Stimulated State
nConcave	184	392
Concave	33	47
Falling	11	25
Rising	5	14
nConvex	1	-
Step-down	-	1

Figure 4.10 - Figure 4.12 illustrate some of these patterns.

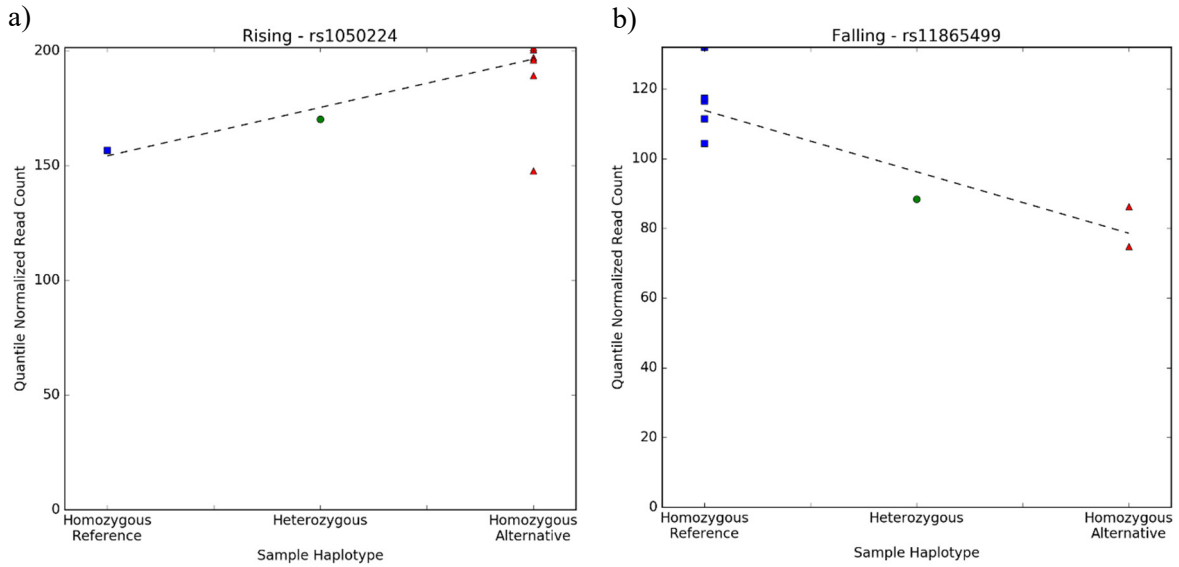


Figure 4.10 – Sample haplotype plots showing the: a) rising and b) falling patterns.

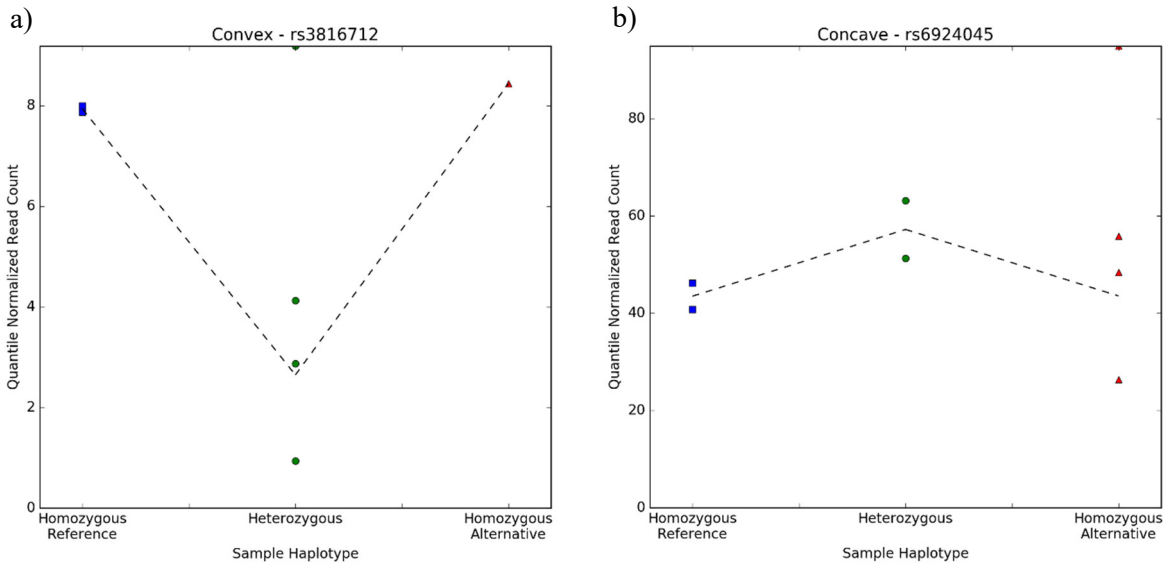


Figure 4.11 - Sample haplotype plots showing the: a) convex and b) concave patterns.

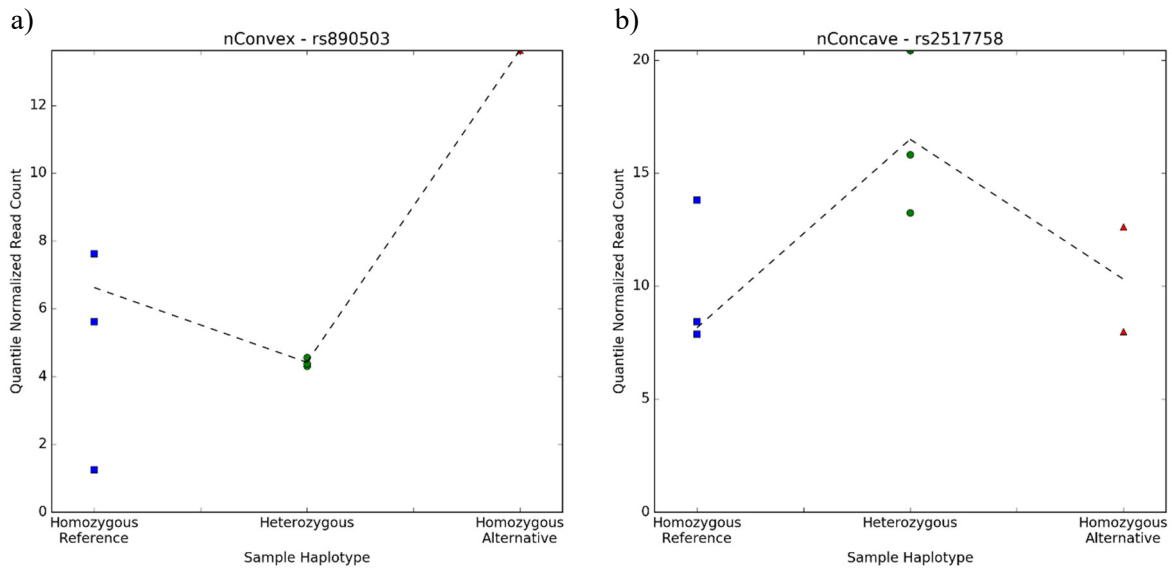


Figure 4.12 - Sample haplotype plots showing the: a) nConvex and b) nConcave patterns.

4.2.2 Linkage Disequilibrium

Table 4.10 depicts the number of intron/upstream variants before and after filtering with respect to linkage disequilibrium.

Table 4.10 – Distribution of variants before and after filtering for a linkage disequilibrium correlation (R^2) of 0.8 or higher, according to category (intron/upstream) and sample state.

	Naive State		Stimulated State	
	Introns	**Upstream**	**Introns**	**Upstream**
No filter	389	72	1003	147
$R^2 \geq 0.80$	154	21	457	63
% Retention	39.59	29.17	45.56	42.86

The upstream variants can also be described in relation to their physical distance to the potentially regulated gene. As a result, Figure 4.13 illustrates the distribution of these base substitutions.

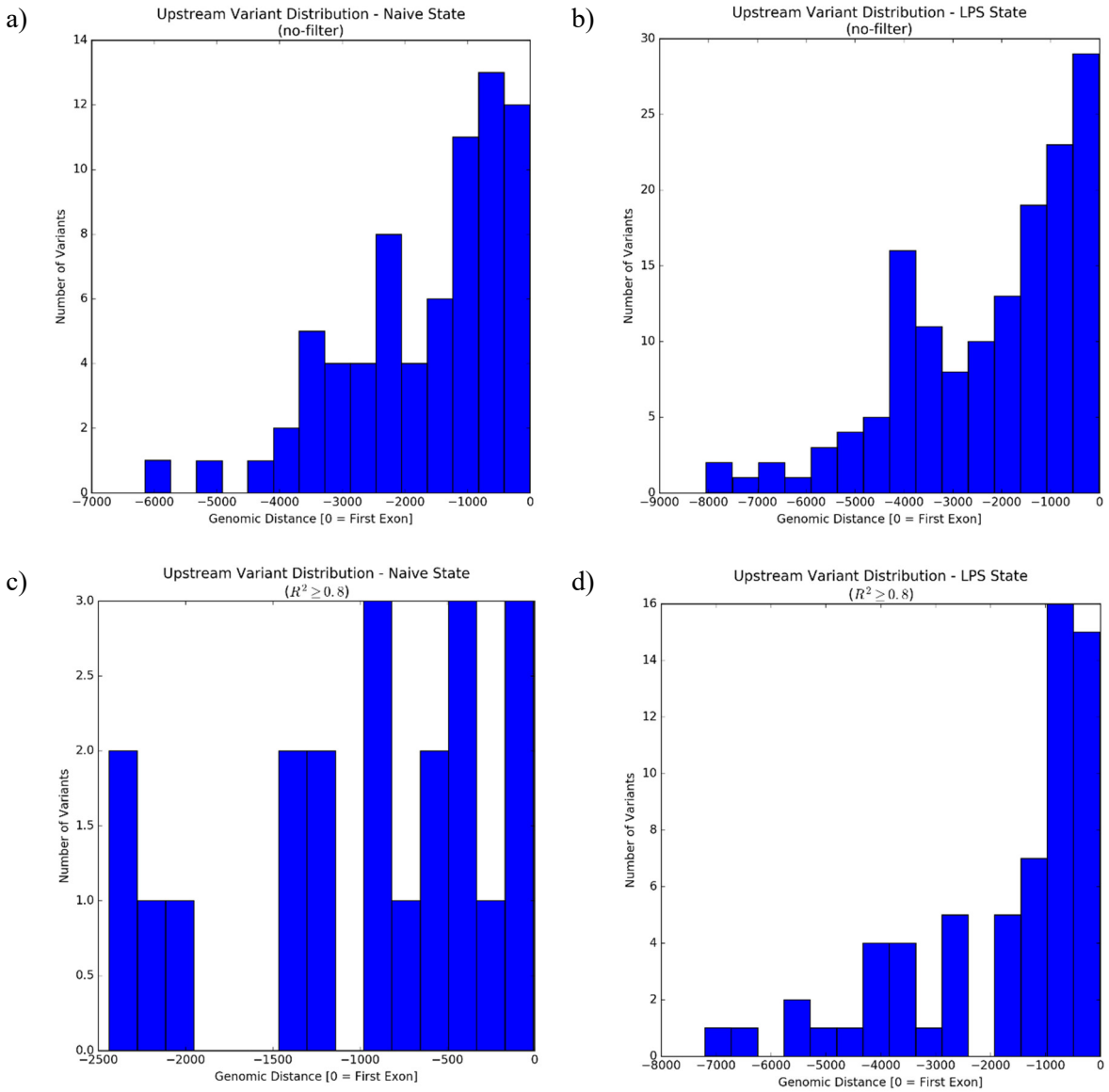


Figure 4.13 – Distribution of the upstream distances between variants and the first exon of the possibly regulated gene, according to state and linkage disequilibrium filter ($R^2 \geq 0.8$): a) basal state and no filter; b) treated state and no filter; c) naïve samples and filtered by LD; and, d) LPS state and filtered by LD.

The selected variants ($R^2 \geq 0.8$) present overlap across states, hence Table 4.11 depicts the number of variants unique to their state as well as the total number of overlapping entries. The genes corresponding to these variants amount to 63 (naïve) and 120 (LPS).

Table 4.11 – Number of potential regulatory variants according to their presence across samples (overlap) or exclusive state membership (uniqueness).

Variant Category	Unique		Overlap	Total	
	Naïve	LPS		Naïve	LPS
Introns	89	392	65	154	457
Upstream	17	59	4	21	63

4.3 Function Evaluation

4.3.1 Variant Annotation

The shortlisted variants were annotated with the Variant Effect Predictor tool from ENSEMBL. Consequently, Table 4.12 presents the base substitutions laying in regions of known regulatory role.

Table 4.12 – Total number of regulatory variants, as annotated by VEP, as well as their percent contribution towards the shortlisted set (variant category and state).

Variant Category	Naïve		Overlap		LPS	
	Unique				Unique	
	Total	Shortlisted Fraction	Total	Shortlisted Fraction	Total	Shortlisted Fraction
Introns	25	28.1%	16	24.6%	113	28.8%
Upstream	6	35.3%	0	0	19	32.2%

Table 4.13 illustrates the data on the selected upstream variants with respect to their distance to the first exon of the potentially regulated gene.

Table 4.13 – Distribution of the upstream variants according to distance and regulatory annotation.

Upstream Distance (bases)	Naïve		Overlap		LPS	
	Unique				Unique	
	Total	Regulatory	Total	Regulatory	Total	Regulatory
0 – 500	5	3	2	0	14	6
500 – 1,000	4	1	2	0	13	4
1,000 – 2,500	8	2	0		12	3
2,500 – 5,000	0		0		16	4
5,000 – 10,000	0		0		4	2
Total	**17**	**6**	**4**	**0**	**59**	**19**

The annotations concerning the regulatory variants span different regulatory elements. Therefore, Figure 4.14 – Figure 4.16 depict pie plots with their composition according to state and variant category.

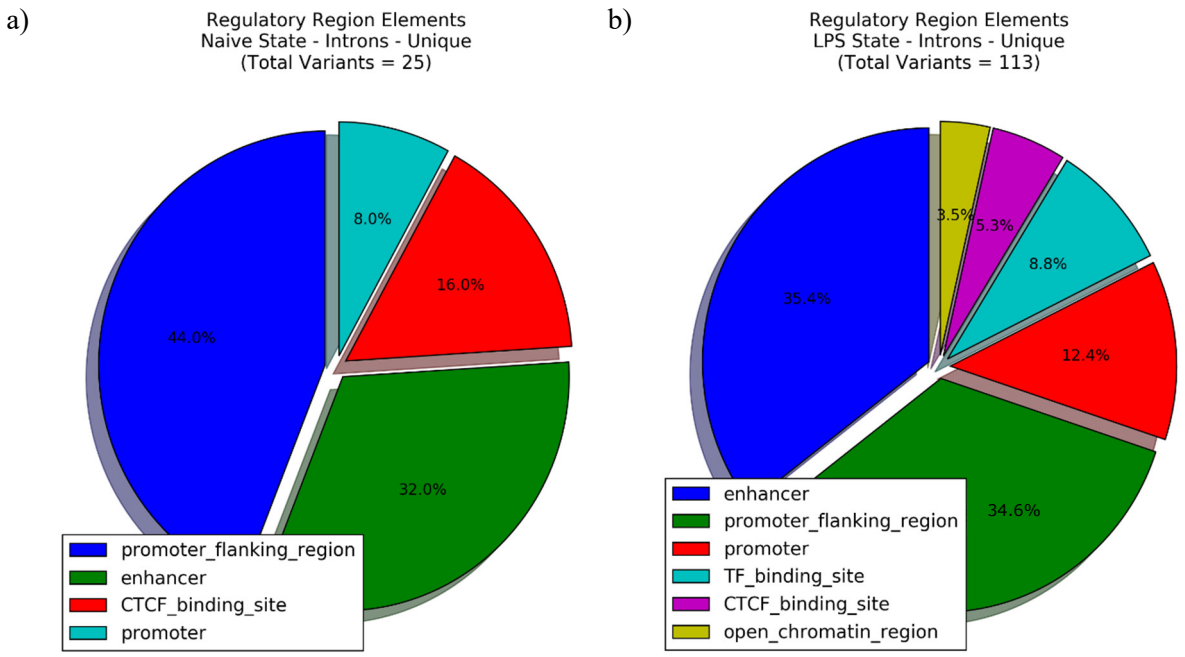


Figure 4.14 – Composition of the unique intronic variants with respect to the type of regulatory element and state: a) naïve, and b) LPS.

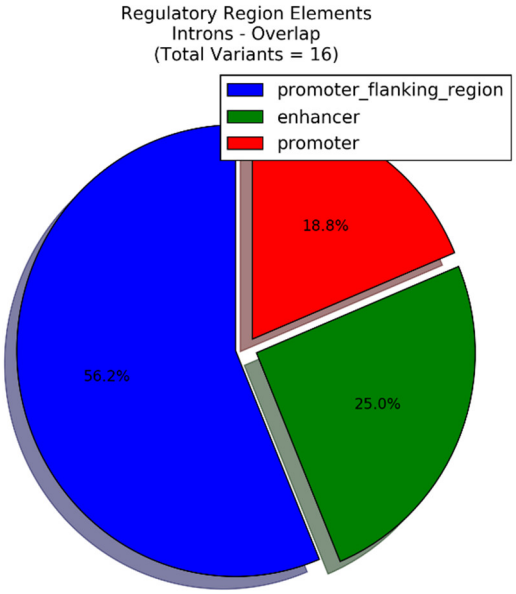


Figure 4.15 – Composition of the overlapping intronic variants with respect to the type of regulatory element.

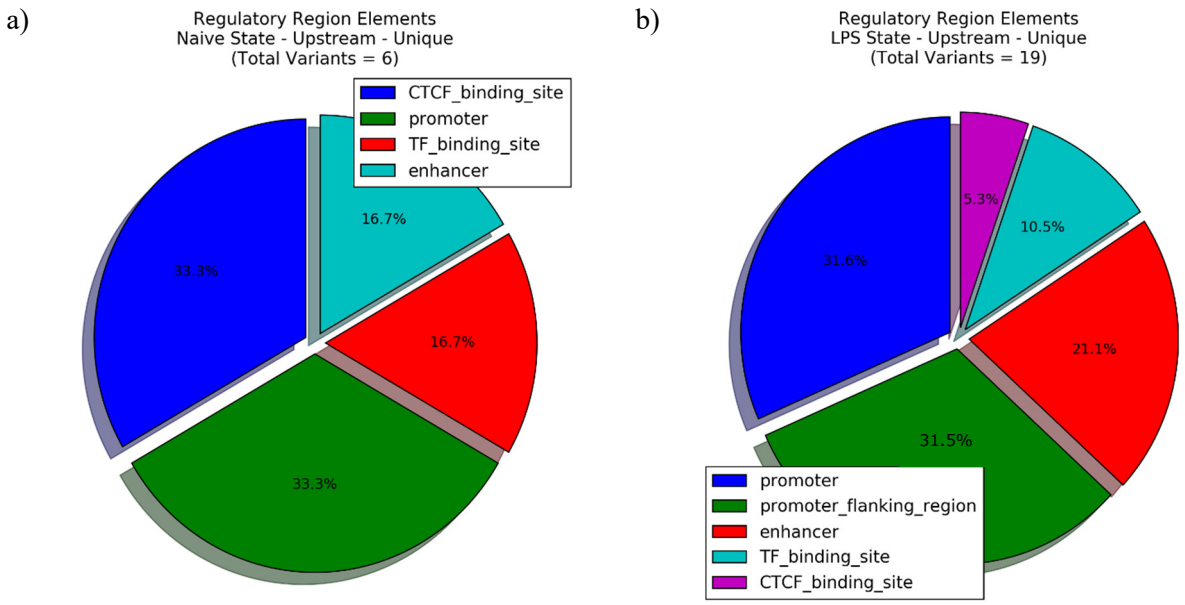


Figure 4.16 – Composition of the unique upstream variants with respect to the type of regulatory element and state: a) naïve, and b) LPS.

4.3.2 Functional Enrichment Analysis

The analysis of the genes containing the selected variants first presents those common to both states, then the unique entries to the LPS samples followed by the naïve scenario. Consequently, Figure 4.17 - Figure 4.19 illustrate the enriched terms in the following gene ontology categories: biological process, molecular function, and pathway. In addition, ToppFun processed 23 of the 25 overlapping genes.

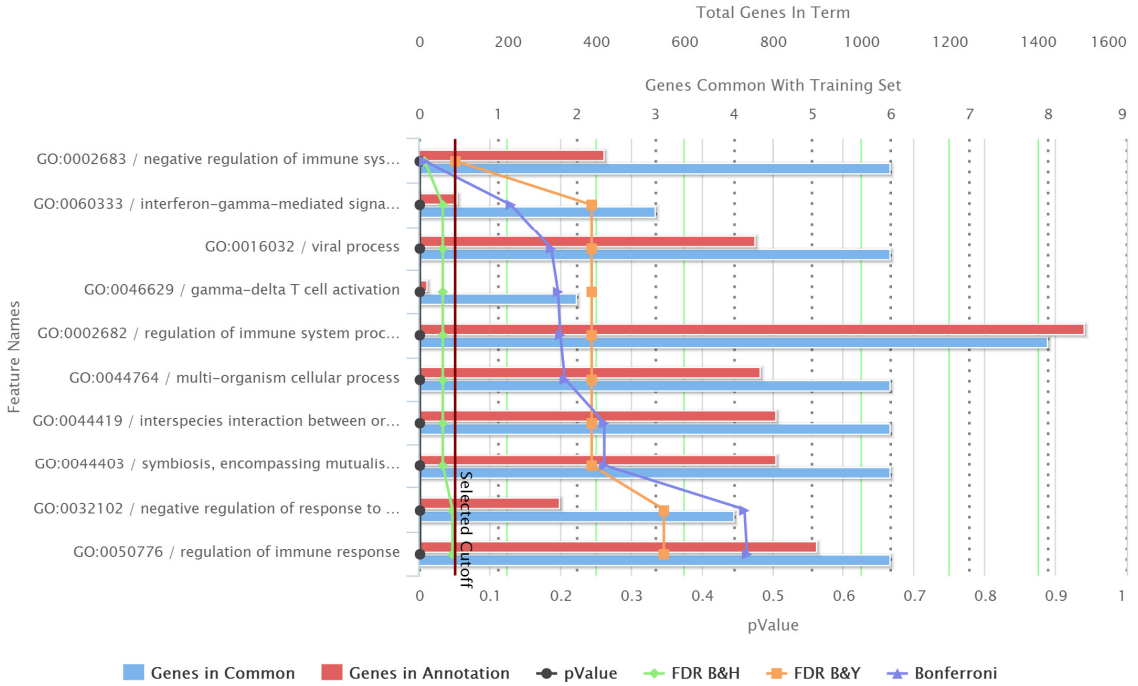


Figure 4.17 – Gene Ontology terms, corresponding to biological processes, enriched in the list of genes present in both states (FDR ≤ 0.05).

Plot generated by ToppFun (Chen et al., 2009).

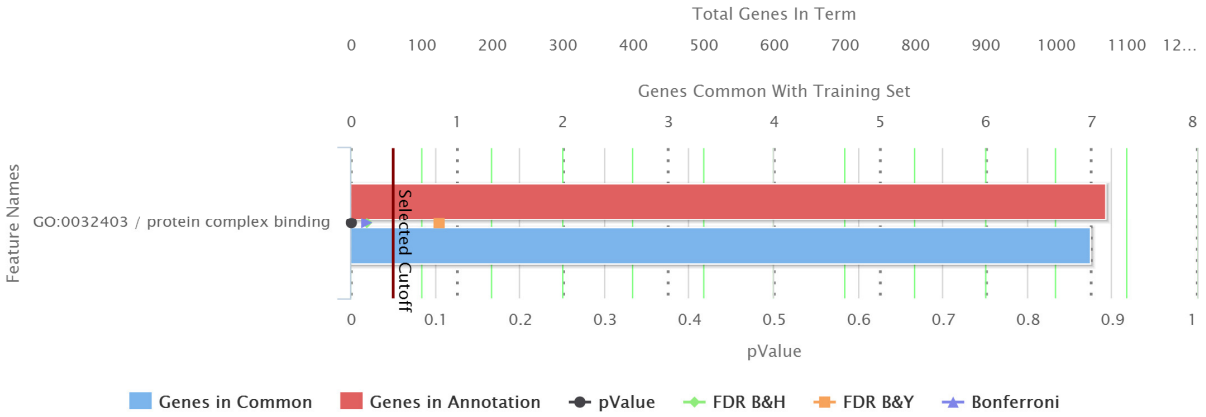


Figure 4.18 – Gene Ontology terms, corresponding to molecular function, enriched in the list of genes present in both states (FDR ≤ 0.05).

Plot generated by ToppFun (Chen et al., 2009).

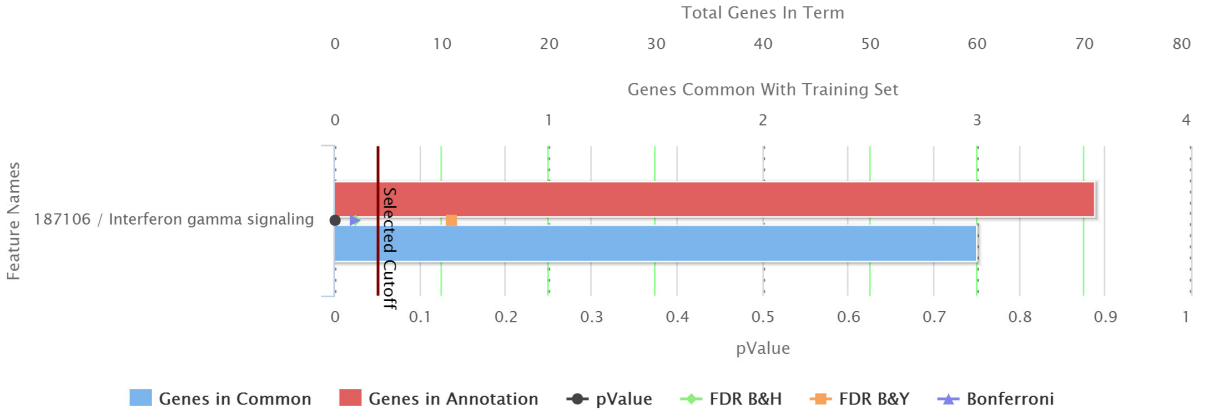


Figure 4.19 – Gene Ontology terms, corresponding to pathway, enriched in the list of genes present in both states (FDR ≤ 0.05).

Plot generated by ToppFun (Chen et al., 2009).

Figure 4.20 and Figure 4.21 illustrate the GO terms concerning biological process and molecular function for the list of genes unique to the LPS state. These plots present a lower FDR threshold only to restrict the number of terms in the graph, since these entries remain the same as those with an FDR of 0.05. In addition, ToppFun retrieved information on 93 of the 95 genes unique to this state.

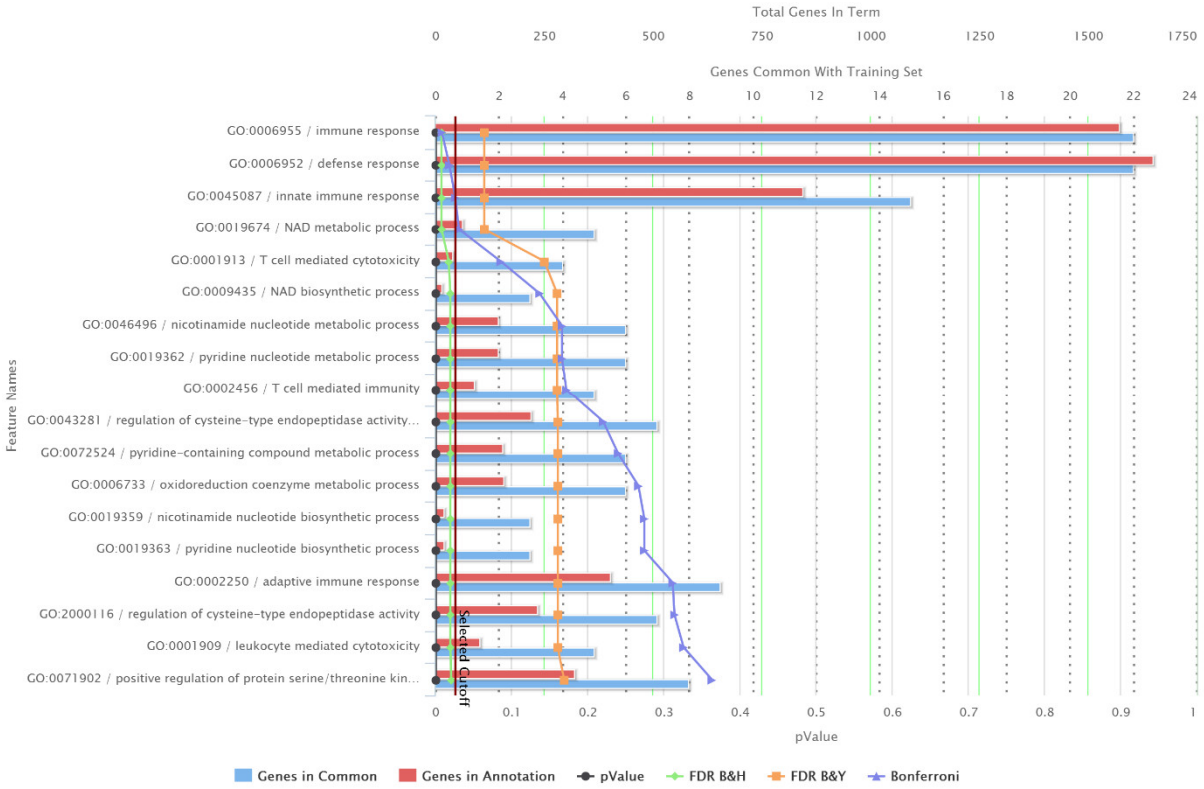


Figure 4.20 – Gene Ontology terms, corresponding to biological processes, enriched in the list of genes unique to the LPS state (FDR ≤ 0.05).

Plot generated by ToppFun (Chen et al., 2009).

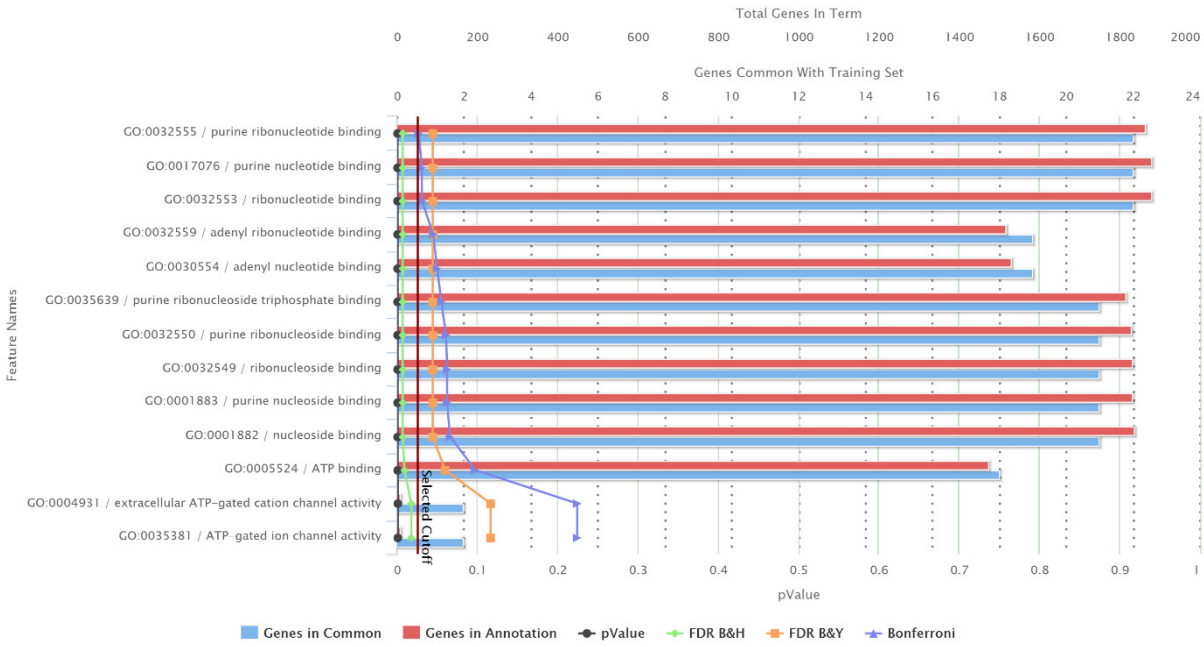


Figure 4.21 – Gene Ontology terms, corresponding to molecular function, enriched in the list of genes unique to the LPS state (FDR ≤ 0.05).

Plot generated by ToppFun (Chen et al., 2009).

The naïve state presented no enrichment terms with an FDR ≤ 0.05 by ToppFun. Furthermore, the g:Profiler tool (Reimand et al., 2007) did not provide any relevant information on the genes of this state.

5 DISCUSSION

5.1 Data Preprocessing

5.1.1 SNP-array

The loss in variants eligible for analysis impacts negatively on the analysis, since it restricts the possibilities to investigate upstream variants. This type of variants has strong potential of regulating gene expression, especially those lying in close proximity to the transcription start site. Nonetheless, the study was not impaired by this scenario and the remaining variants were successfully employed in downstream analysis.

5.1.2 RNA-sequencing

5.1.2.1 Preprocessing

The original data covered male and female samples as well as varying sequencing depths. In addition, the analysis divided the reads according to the flow cell lanes of the sequencing machine. Therefore, Table 4.1 shows varied outcomes on the approaches employed to retain high quality reads. Overall, the strategy reduced the number of available sequences. However, the remaining reads are enriched by high quality nucleotides, thus improving the reliability of the variant calling procedure. The highest reduction occurred on sample 1 of the naïve state (33.24%) and sample 9 of the LPS state (32.22%). On the contrary, the samples with the highest quality in the original data were number 6 for the unstimulated state and number 4 for the treated scenario, since these samples retained 91.77% of the original reads.

This analysis explores in more detail sample 1 of the untreated scenario and sample 2 of the stimulated state. However, these findings can be extended, at varying degrees, to all samples within their respective treatment groups. The choice of these particular samples was motivated by the level of reduction in available reads as well as by peculiarities concerning some flow cell lanes (2 and 4) of the sequencing machine.

Figure 4.1a depicts sample 1 and lane 1 (S1L1) and it represents a typical sequencing run. The read shows gradual improvement in quality at the calibration cycles, a steady performance of the sequencing machine towards the middle of the sequence and a more pronounced loss in efficiency towards the end. In fact, sample 2 and lane 2 (S2L2) also show the same general trends (Figure 4.2a). However, it also presents a systematic error at the initial runs (bases 3-4). This behavior could be addressed by trimming the reads from these lanes, so the resulting sequence becomes shorter. Nonetheless, such measure was not taken, because the variant calling procedure already relies on

recalibrating the sequencing quality. Therefore, it should be able to discard the problematic bases without interfering with the surrounding nucleotides, which exhibit excellent quality.

The GC content distribution informs on the need for ribosomal contamination, since RNA from this source is enriched on these nucleotides. In fact, Figure 4.1b and Figure 4.2b demonstrate small deviation in the shape of the empirical distribution, with respect to the theoretical, towards the region with higher GC content. However, the contaminant filter showed no improvement on the data (Figure 4.3b and Figure 4.4b). In other words, the experimental procedure was able to greatly remove this source of error and this computational strategy was unable to further improve it. Therefore, these files did not continue down the pipeline. This decision was motivated by the presence of scaffolds in the reference assembly, since these regions already contain regions enriched in highly expressed RNA sequences from human ribosomes. Thus, the mapping procedure alone might remove remaining inappropriate sequences. In addition, the contaminant removal showed no effect on the overall quality of the reads, since Figure 4.3a and Figure 4.4a resemble their counter parts in the original data (Figure 4.1a and Figure 4.2a).

Trimming led to changes both in GC content and read quality. This operation generated a slight increase in the GC content of the remaining reads, as depicted in Figure 4.5b and Figure 4.6b. In addition, it provided considerable improvement in sequence quality, since the remaining reads become enriched for those exhibiting high score. The beneficial aspects of this operation outweighs the loss in numbers, because it decreases downstream processing time and increases the reliability of the calls from the variation analysis. The stark difference is noticeable in Figure 4.5a and Figure 4.6a, since all bases are above a Phred score of 28 and considerably above the minimum for this study (20). The bases 3-4 on S2L2 (Figure 4.6b) correspond to sequencing artifacts, hence they remain on file for treatment by the GATK variant calling pipeline.

The preprocessing stage evaluated the contamination of the samples by ribosomal RNA and processed quality control measures in the sample files. There is no evidence of contamination, since the filtration yielded marginal decrease in reads with elevated GC content. At the same time, trimming the reads was successful in enriching the files for high quality sequences.

5.1.2.2 Alignment

The use of STAR increases the confidence of the alignment procedure, since it borrows information from a splice junction database as well as from an initial temporary alignment. Consequently, the two-step process increases the sensitivity of the variant calling pipeline.

5.1.2.3 Variant Calling

The estimation of the genetic variation consists of a challenging task, especially if the evaluation comes from the analysis of RNA instead of DNA. Therefore, the previous measures aimed at providing this pipeline with reliable and high quality data. However, a faithful depiction of the

biological samples also influences this procedure, hence the use of the indel realignment and base quality score recalibration tools.

The first algorithm evaluates the edges of splice junctions for possible misaligned segments. If such regions are found, then it realigns them in a *de novo* manner. This method improves the accuracy in determining the correct read assignment for these regions. Thus, it decreases the rate of false positive variant calls. These variants play an important role on the GeneiASE software, since it makes no distinction between indels and base substitutions. Moreover, it shortlists the genes exhibiting strong potential of exhibiting ASE for the evaluation of nearby variants with respect to their regulatory role. Consequently, a faithful variant estimation includes indel regions and this outcome directly impacts the identity of the genes selected for analysis.

The second tool decreases the bias introduced either by the aligner or by the sequencing machine. For instance, Figure 4.7a and Figure 4.8a illustrate an overly confident sequencing machine, since the reported quality score is often higher than the empirical score before recalibration. However, after the procedure this trend was eliminated, hence the two scores follow more closely the y = x line of the plot. This correction directly impacts the proposed model, since it heavily relies on base substitutions, especially to determine if a sample is heterozygous or homozygous with respect to the alleles of the variant (reference/alternative). Therefore, the correct base quality score influences the variant identity as well as the allele read counts.

This correction can also be observed by analyzing the base distribution according to quality score, as depicted in Figure 4.7b and Figure 4.8b. The nucleotide frequency is more spread along the range of quality scores and clearly form a unimodal distribution instead of a bimodal enriched around near perfect scores.

These measures contributed to an excellent performance of the haplotype caller across samples and states. The calls from the naïve samples ranged from 166k to 313k (Table 4.2) while the stimulated samples elicited from 213k to 513k calls (Table 4.3). Moreover, the quality filter reduced these numbers in approximately 11 – 19%. Nonetheless, this analysis relies on base substitutions, which led to a further decrease of around 21 – 27%, with respect to the high quality set. Lastly, the final set corresponds to 13 – 30% of the unfiltered variant calls. These mutations contain representatives in all three haplotype configurations as well as 10 reads across alleles in heterozygous individuals.

The overall number of variants consists of the union of the sets from the individual subjects. Therefore, some variants might show less than eight individuals but never less than three, which is the minimum number to generate a haplotype plot.

In summary, the variant calling pipeline successfully detected base substitutions and indels. It increased the confidence of variants around indels by realigning the reads spanning splice junctions, and it corrected the read quality by evaluating the empirical and reported base scores.

5.1.2.4 Selection Criteria for ASE Genes

The ability of the model to shortlist variants depends on the level of confidence given by GeneiASE (Edsgärd et al., 2016). Therefore, this study evaluated different combinations of FDR thresholds and individuals simultaneously exhibiting the selected gene (Table 4.4 and Table 4.5).

The general trend of these associations shows that as FDR increases so does the number of shortlisted genes. However, an increase in the number of individuals implies further restriction, thus it leads to a decrease in the number of selected genes. Clearly, there must be a compromise between the two parameters, in order to retain an adequate and faithful level of information for downstream analysis. Consequently, the chosen parameters consist of an upper FDR limit of 0.05, to control the number of false positives, while the number of samples corresponds to the minimum required to generate a haplotype plot.

This combination of factors leads to 535 (naïve) and 884 (LPS) eligible genes. Furthermore, there is a noticeable difference between the two states. The unstimulated state shows 144 unique genes while the treated samples contain 493, Table 4.4 and Table 4.5, respectively. This trend is in agreement with the number of unique variants found in Table 4.2 and Table 4.3, since they depict 10,367 base substitutions for the naïve scenario and 31,094 for the LPS treated samples.

The difference between states illustrates the extend of the immune response triggered by the treatment with lipopolysaccharide, since it elicited the expression of a greater number of genes than in the basal state. In addition, both the unique and the overlapping genes/variants are appreciated in the analysis.

The following step comprises the concatenation of the synthetic gene boundaries as well as the data from the SNP-array and RNA-sequencing. The synthetic boundaries consist of a conservative measure to avoid the overlap of intron/upstream variants with exons from different isoforms. Therefore, all intronic/upstream mutations lie in *loci* that could harbor regulatory elements. In addition, both data sources provide variants for the intron and upstream categories while only RNA-sequencing contributes to exons, due to the need of read counts for shortlisting exon variants. In fact, Table 4.6 presents the total number of variants according to the categories of the study (exon, intron, and upstream) as well as biological sample condition (basal and stimulated states). In essence, the LPS treated samples have on average twice as much variants than the naïve state.

This study evaluated different combinations of factors for selecting genes with potential ASE. Firstly, it established an upper limit on the rate of false positives by setting the FDR threshold to 0.05. Secondly, it chose the minimum number of individuals to support the generation of haplotype shapes. Consequently, these parameters led the number of variants for the treated samples to be approximately double of that for the basal state.

5.1.2.5 Quantile Normalization

The quantile normalization transformed the discrete read count in a continuous variable. Therefore, the number of options for fitting the data greatly expanded. At the same time, the data shows some peculiar characteristics, such as: it only takes positive values; it shows a sharp rise close to zero followed by a very long right tail.

These traits place the exponential, lognormal and Weibull distributions as prime candidates. The exponential distribution satisfies the positive criterion, but it does not describe well the region of low read counts. In fact, it over emphasizes the contribution of this region. The lognormal distribution also shows difficulties with this region, since it is unable to cope with the steep rise which leads to a delayed response with a fat right tail. Finally, the inverted Weibull distribution agrees very well with the data, since it is very dynamic in handling the fast rise and somewhat slow decay.

The data from both states shows similar trends. Therefore, each sample set was fitted by an inverted Weibull distribution. Table 4.7 depict the fit parameters while Figure 4.9 illustrates the agreement between the fit and the data.

5.2 Shortlist Variants

5.2.1 Custom Model

The main goals of the model are to shortlist exon variants and provide exon-intron/upstream variant pairs for linkage disequilibrium analysis. Therefore, it generates haplotype plots of the exon variants and compares the test statistic from the biological data against a null model. Therefore, Table 4.8 presents the number of occurrences from each pattern according to state while Table 4.9 depicts the shapes from the shortlisted variants.

The ranks of the most occurring patterns coincide across states. In other words, the non-symmetrical V-shapes are the most abundant haplotype-plots (nconcave and nconvex), followed by the linear-shapes (falling and rising) and the symmetrical V-shapes (concave and convex), as depicted in Table 4.8. The inconclusive pattern category consists of plots with no clear pattern, therefore such occurrences are discarded. Finally, the flat pattern amounted to 74 and 79 plots for the naïve and stimulated states, respectively. Therefore, approximately 78% (basal state) and 81% (stimulated state) of the variants seem to represent ASE genes or exhibit a discernible pattern supporting this assumption.

The test statistic enables the shortlisting process, hence an FDR of 0.05, resulted in 234 exon variants in the naïve state and 740 on the treated samples. Moreover, Table 4.9 depicts the count of the surviving patterns. In essence, only the nconcave and concave remained as the most abundant shapes both across states and within the V-shape category. The intermediate positions were occupied by the falling and rising patterns while the last place differed between the naïve and the treated samples. In fact, the shape with the lowest occurrence in the first sample group is the nconvex while step-down closes the list for the second.

Figure 4.10 through Figure 4.12 provide a visual interpretation of these numbers, since they illustrate the haplotype plots of the shortlisted variants (FDR ≤ 0.05). Furthermore, they illustrate how well the model handles outliers. For instance, Figure 4.10a depicts the rising pattern of variant rs1050224. This plot presents an outlier in the homozygous to the alternative allele individuals, thus the analysis disregards this data point. Moreover, Figure 4.11b and Figure 4.12a show similar behavior for the subjects homozygous to the alternative (rs6924045) and to the reference allele (rs890503).

This strategy compares the expression level of the alternative allele of heterozygous individuals to that of individuals carrying the homozygous haplotypes. Therefore, the generation of the test statistic presents a meaningful measure against which to compare a null distribution. This characteristic proves superior to a simple comparison of read counts from heterozygous individuals and random samples of the quantile normalized distribution. Furthermore, this method prevents bias towards highly expressed genes by taking the majority vote on the pattern representing the haplotype plot. In essence, this strategy selects exon variants among ASE genes that are representative of this phenomenon in their respective genes.

5.2.2 Linkage Disequilibrium

Pairs of variants that show strong signs of being inherited together could also inform on the appearance of particular phenotypes or regulatory role. Therefore, this study selected exon variants to pair them with intron and upstream variants. At the same time, the choice of variant pairs was restricted to genes exhibiting allele specific expression. Consequently, the final variant selection informs on the potential of the upstream and intronic base substitutions to regulate the expression of genes with allelic imbalance.

The pairing occurred between exon/intron variants as well as exon/upstream variants within a genomic distance of up to 10,000 bases upstream of the first exon of the given gene. In addition, the pairing does not differentiate between up or down regulation. This characteristic could have been further explored, since the exon selection evaluates the expression from all alleles as well as their deviations across haplotypes. Nonetheless, pairs presenting a linkage disequilibrium correlation greater than or equal to 0.80 were selected as potential regulatory variants.

The filter greatly reduced the number of variants, as depicted in Table 4.10. This table presents the number of evaluated pairs as well as the amount of retained intronic and upstream variants. Overall, the retention levels varied from approximately 29% - 45%. In addition, the naïve state suffered the greatest reductions.

This behavior can be visually interpreted in the upstream variant category by analyzing the change in the shape of the histograms depicting the filtering procedure with respect to genomic distance (Figure 4.13). In fact, the LPS state shows enrichment in the proximal distances both before and after selection while the naïve state only shows this behavior before filtering. The untreated samples retained only 21 upstream variants and they are roughly uniformly distributed.

Another interesting trait of the data is the immune response elicited in the treated samples. This mechanism triggered the expression of genes that are not present in the basal state. In addition, the opposite has also been found to be true, namely, the basal state also contains genes that are unique to its state. Nonetheless, this uniqueness might also be an artifact from the selection procedure, since the retained variants must contain at least 10 read counts across alleles on individuals showing the heterozygous haplotypes. Biologically, the cell might have deactivated some programs to allocate resources to fight the threat of a bacterial infection, since lipopolysaccharide consists of a typical membrane component of Gram-negative bacteria.

Table 4.11 depicts the number of variants unique to each state as well as those placed in genes expressed in both situations. The LPS state has the greatest numbers of unique variants both for intronic and upstream variants. Moreover, there is a greater overlap of intronic variants than of upstream base substitutions.

The variants obtained by this strategy eliminates passing mutations, since retained variants show a significant measure of common inheritance in the European population. Therefore, this enrichment applied to the analysis of genes exhibiting ASE rendered a pool of variants with potential regulatory role.

5.3 Function Evaluation

5.3.1 Variant Annotation

The validation of these findings would rely on the experimental investigation of the selected variants. However, variant annotation provides the means to further restrict the list of potential variants as well as to evaluate the model performance. Consequently, this study employed the VEP annotation tool from ENSEMBL.

The model successfully selected variants with known regulatory role (Table 4.12), since a portion of the chosen mutations had their annotations linked to regulatory elements. The highest number of variants affecting regulatory elements occurred in the intron category with 113 hits in the LPS state and 25 in the naïve samples. The upstream region had lower absolute numbers but it exhibited the highest proportion of correctly selected regulatory variants. Its fractions correspond to 35.3% and 32.2%, for the naïve and LPS states, respectively. In essence, the model provided accurate predictions in a range of 24 – 35% of the cases.

The upstream variants are spread across a region spanning 10 kb towards the 5’ end of the gene. Therefore, Table 4.13 depicts the number of occurrences in terms of the distance from the variant to the first exon of the affected gene. This table shows that correctly assigned regulatory variants do not preferentially occur at a specific position. Nonetheless, the data on upstream regions is very scarce and trends are not easily distinguishable. This behavior could have been strongly influenced by the loss of genotype calls from the SNP-array, since the upstream variants are for their most part heavily dependent on this information.

The use of the synthetic splice junctions decreased computational cost on classifying intronic, exonic and upstream variants, since it collapsed the splice junctions of different isoforms into a single set. However, the use of an external annotation tool provided the effect of the variants in their original setting. Therefore, the types of regulatory elements depicted in Figure 4.14 and Figure 4.15 lie at the core promoter region or in proximal components. This result implies that some variants catalogued in the model as intronic also show evidence of laying in the upstream region. In fact, the intronic category is rich in promoter elements, enhancers, transcription factor sites and the CTCF repressor binding site.

The upstream variants unique to the naïve state show four activator elements and two repressors while those mutations unique to the LPS samples are in their majority activators, as depicted in Figure 4.16. In addition, the model selected only four variants in the overlapping region of the two states, but none of these mutations have been annotated to regulatory regions.

The model could increase its potential in selecting variants by revising the assumptions on the generation of the synthetic gene boundaries. The current approach merges the exons across isoforms, hence the model does not probe some intronic/upstream regions. Therefore, a different strategy could be to use the full set of splice junctions or to take the intersection of the exons instead of the union. The latter corresponds to classify an intronic/upstream variant as such if at least one isoform supports this classification. This measure would generate variant catalogues that best represent the biological evidence.

In summary, the current approach was successful in proposing variants with potential regulatory role, since the annotation of the selected variants showed regulatory features in 24 – 35% of the shortlisted variants. Nonetheless, the model could improve on the classification of variants according to their genomic position as well as by a greater number of genotype calls covering regions not transcribed.

5.3.2 Functional Enrichment Analysis

The functional enrichment analysis provided annotations that are enriched by the genes selected in this study. These terms were restricted to biological processes and molecular functions. The first illustrates the purpose of the gene activity, in other words, it represents the biological objective while the second depicts the biochemical reaction. In addition, there was no background subtraction, since the analysis no longer is selective but descriptive.

The overlapping genes represent terms describing the immune system (Figure 4.17 - Figure 4.19). In fact, the four most significant terms in Figure 4.17 involve the immunological machinery. Furthermore, the interferon gamma signaling annotation places very high in this list and is the sole metabolic pathway detected by this analysis (Figure 4.19). This protein is a master regulator of many cellular programs related to immune response (Larkin, Ahmed, Wilson, & Johnson, 2013). Therefore, the data corresponds to the typical genes expected from white blood cells.

The treated samples follow similar rationale, since the cell response evokes regulators of the immune system. Figure 4.20 lists immune response, defense response and innate immune response as the most significant terms. This behavior was expected, since lipopolysaccharide simulates a bacterial attack. Consequently, the cell has triggered its mechanisms to defend the body against the antigen. This picture becomes even more pronounced by considering the molecular functions (Figure 4.21), since these are in their majority related to nucleotide metabolism. Finally, the naïve state contained genes without known annotations even when tested by two different software.

The terms associated with the selected genes represent quite well the anticipated behavior from white blood cells. Especially, those terms illustrating the response of the cell against a bacterial threat.

6 CONCLUSION

The samples showed little to no ribosomal RNA contamination and trimming the reads greatly improved the overall read quality. Furthermore, variant calling was processed with information on flow cell lane, in order to remedy sequencing artifacts found in lanes 2 and 4. This pipeline also benefited from indel realignment and base quality score recalibration to improve the confidence around splice junctions and to raise the reliability of the sequencing scores. As a result, the variant calling procedure provided an abundance of polymorphisms.

The approach to process the gene boundaries could be revised to better suit external sources or to analyze ASE at an isoform level. Therefore, the model could benefit from evaluating the performance of a synthetic construct based on the intersection of exons rather than the union. Also at a gene level, the chosen parameters for the GeneiASE software supplied the analysis with suitable gene candidates.

Quantile normalization proved a feasible methodology to compare the data across samples. Furthermore, it provided a common distribution from which to sample read counts to generate variant null model. The data for both scenarios agrees very well with the inverted Weibull distribution.

Approximately 78% (basal state) and 81% (stimulated state) of the exon variants seem to represent ASE genes or exhibit a discernible pattern supporting this assumption. Furthermore, the use of linkage disequilibrium to select base substitutions with regulatory potential eliminates passing mutations and retains those with high chance of co-segregation.

The model successfully selected variants known to play a role in gene regulation. Furthermore, the genes possibly regulated by them exhibit typical programs of white blood cells. Consequently, this methodology could greatly benefit from an increased sample size as well as an enlarged genotype dataset on variants spanning not transcribed DNA regions.

REFERENCES

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212. http://doi.org/10.1038/nrg3891

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., … Walter, P. (2014). *Essential Cell Biology* (4th Editio). New York, NY: Garland Science, Taylor & Francis Group.

Alon, U. (2007). An Introduction to Systems Biology: Design Principles of Biological Circuits. *Chapman HallCRC Mathematical and Computational Biology Series*. http://doi.org/citeulike-article-id:1314150

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. http://doi.org/10.1038/nature15393

Auwera, G. A. Van Der, Carneiro, M. O., Hartl, C., Poplin, R., Levy-moonshine, A., Jordan, T., … Depristo, M. A. (2014). *From FastQ data to high confidence varant calls: the Genonme Analysis Toolkit best practices pipeline*. *Curr Protoc Bioinformatics* (Vol. 11). http://doi.org/10.1002/0471250953.bi1110s43.From

Beadle, G. W., & Tatum, E. L. (1941). Genetic Control of Biochemical Reactions in Neurospora Author (s): G . W . Beadle and E . L . Tatum Source : Proceedings of the National Academy of Sciences of the United States of America , Published by : National Academy of Sciences Stable URL : http: *Proceedings of the National Academy of Sciences of the Uniteds States of America1*, *27*(11), 499–506.

Bolstad, B. M., Irizarry, R. ., Astrand, M., & Speed, T. P. (2003). Reading 5 (1) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*(2), 185–193. http://doi.org/10.1093/bioinformatics/19.2.185

Branden, C., & Tooze, J. (1999). *Introduction to Protein Structure* (2nd Editio). Garland Science.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., … Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, *42*(D1), 471–480. http://doi.org/10.1093/nar/gkt1103

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. http://doi.org/10.1186/s13742-015-0047-8

Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(SUPPL. 2), 305–311. http://doi.org/10.1093/nar/gkp427

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. http://doi.org/10.4161/fly.19695

Clark, D. P., & Pazdernik, N. J. (2012). Biotechnology. In *Elsever Inc.* (p. 750).

Crick, F. H. C. (1970). Central Dogma of Molecular Biology. *Nature*. http://doi.org/10.1038/227561a0

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., … D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42*(D1), 481–487. http://doi.org/10.1093/nar/gkt1102

Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F., & Marchini, J. (2013). Haplotype estimation using sequencing reads. *American Journal of Human Genetics*, *93*(4), 687–696. http://doi.org/10.1016/j.ajhg.2013.09.002

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8. http://doi.org/10.1038/ng.806

Edsgärd, D., Iglesias, M. J., Reilly, S.-J., Hamsten, A., Tornvall, P., Odeberg, J., & Emanuelsson, O. (2016). GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Scientific Reports*, *6*(October 2015), 21134. http://doi.org/10.1038/srep21134

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., … Bertone, P. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, *10*(12), 1185–91. http://doi.org/10.1038/nmeth.2722

Fourel, G., Magdinier, F., & Gilson, É.́. (2004). Insulator dynamics and the setting of chromatin domains. *BioEssays*, *26*(5), 523–532. http://doi.org/10.1002/bies.20028

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., … Wishart, D. S. (2009). SMPDB: The small molecule pathway database. *Nucleic Acids Research*, *38*(SUPPL.1), 480–487. http://doi.org/10.1093/nar/gkp1002

Fulton, D., Sundararajan, S., Badis, G., Hughes, T., Wasserman, W., Roach, J., & Sladek, R. (2009). TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol*, *10*(3), R29. http://doi.org/10.1186/gb-2009-10-3-r29

Gaffney, D. J. (2013). Global Properties and Functional Complexity of Human Gene Regulatory Variation. *PLoS Genetics*, *9*(5), 1–8. http://doi.org/10.1371/journal.pgen.1003501

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*, *8*(6), 469–477. http://doi.org/10.1038/nmeth.1613

Gershenzon, N. I., & Ioshikhes, I. P. (2005). Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, *21*(8), 1295–1300. http://doi.org/10.1093/bioinformatics/bti172

Gregory, T. R. (2005). Synergy between sequence and size in Large-scale genomics. *Nature Reviews Genetics*, *6*(9), 699–708. http://doi.org/10.1038/nrg1674

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. R. (2003). Rfam: An RNA family database. *Nucleic Acids Research*, *31*(1), 439–441. http://doi.org/10.1093/nar/gkg006

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. http://doi.org/10.1093/nar/gkn923

Jacob, F., & Monod, J. (1961). *On the Regulation of Gene Activity*. *Cold Spring Harb Symp Quant Biol* (Vol. 26). ACADEMIC PRESS, INC. http://doi.org/10.1101/SQB.1961.026.01.024

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression : how the genome integrates intrinsic and environmental signals, *33*(march), 245–254. http://doi.org/10.1038/ng1089

Jewison, T., Su, Y., Disfany, F. M., Liang, Y., Knox, C., MacIejewski, A., … Wishart, D. S. (2014). SMPDB 2.0: Big improvements to the small molecule pathway database. *Nucleic Acids Research*, *42*(D1), 478–484. http://doi.org/10.1093/nar/gkt1067

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., & De Bakker, P. I. W. (2008). SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap.

Bioinformatics, *24*(24), 2938–2939. http://doi.org/10.1093/bioinformatics/btn564

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. http://doi.org/10.1093/nar/gkv1070

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. http://doi.org/10.1093/bioinformatics/bts611

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., … Pico, A. R. (2015). WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, *44*(October 2015), gkv1024. http://doi.org/10.1093/nar/gkv1024

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research*, *37*(13), 4181–4193. http://doi.org/10.1093/nar/gkp552

Laird, N. M., & Lange, C. (2011). *The Fundamentals of Modern Statistical Genetics*. (M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, W. Wong, & For, Eds.) (Vol. 1). New York, NY: Springer Science+Business Media, LLC. http://doi.org/10.1007/978-1-4419-7338-2

Larkin, J., Ahmed, C. M., Wilson, T. D., & Johnson, H. M. (2013). Regulation of interferon gamma signaling by suppressors of cytokine signaling and regulatory T cells. *Frontiers in Immunology*, *4*(DEC), 1–8. http://doi.org/10.3389/fimmu.2013.00469

Larson, D., & Abbott, T. (2016). Bam-Readcount. Retrieved May 21, 2016, from https://github.com/genome/bam-readcount

Lefebvre, J. F., Vello, E., Ge, B., Montgomery, S. B., Dermitzakis, E. T., Pastinen, T., & Labuda, D. (2012). Genotype-based test in mapping Cis-regulatory variants from Allele-specific expression data. *PLoS ONE*, *7*(6). http://doi.org/10.1371/journal.pone.0038667

Levine, M., & Tijan, R. (2003). Transcription regulation and animal diversity. *Nature*, *424*, 147–151.

Mannervik, M., Nibu, Y., Zhang, H., & Levine, M. (1999). Transcriptional coregulators in development. *Science*, *284*(5414), 606–609. http://doi.org/10.1126/science.284.5414.606

Maston, G. A. et al. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, *7*(1), 29–59. http://doi.org/10.1146/annurev.genom.7.080505.115623

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*(16), 2069–2070. http://doi.org/10.1093/bioinformatics/btq330

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, *16*(9), 1182–1190. http://doi.org/10.1101/gr.4565806

Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics*, *5*(4), 355–364. http://doi.org/10.1093/bib/5.4.355

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *27*(1), 29–34. http://doi.org/10.1093/nar/27.1.29

Phipson, B., & Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, *9*(1), Article39. http://doi.org/10.2202/1544-6115.1585

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. http://doi.org/10.1093/nar/gks1219

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). G:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, *35*(SUPPL.2), 193–200. http://doi.org/10.1093/nar/gkm226

Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., … Pico, A. R. (2007). GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, *8*, 217. http://doi.org/10.1186/1471-2105-8-217

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., & Buetow, K. H. (2009). PID: The pathway interaction database. *Nucleic Acids Research*, *37*(SUPPL. 1), 674–679. http://doi.org/10.1093/nar/gkn653

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–11. http://doi.org/10.1093/nar/29.1.308

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445. http://doi.org/10.1073/pnas.1530509100

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a, … Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–50. http://doi.org/10.1073/pnas.0506580102

Sun, W. (2012). A Statistical Framework for eQTL Mapping Using RNA-seq Data. *Biometrics*, *68*(1), 1–11. http://doi.org/10.1111/j.1541-0420.2011.01654.x

The BROAD Institute. (2016). Picard Tools. Retrieved April 20, 2016, from http://broadinstitute.github.io/picard/

The ENCODE Project Consortium, Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a, … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247

The Gene Ontology Consortium. (2000). Gene ontologie: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. http://doi.org/10.1038/75556.Gene

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., … Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141. http://doi.org/10.1101/gr.772403

University of Waikato. (2011). Cell, chromosomes and DNA. Retrieved from http://sciencelearn.org.nz/Contexts/Uniquely-Me/Sci-Media/Images/Cell-chromosomes-and-DNA

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, *12*(11), 1061–3. http://doi.org/10.1038/nmeth.3582

Wadi, L., Meyer, M., Weiser, J., Stein, L., & Reimand, J. (2016). Impact of knowledge accumulation on pathway enrichment analysis. *bioarXiv*, *29*(February), 1–13. http://doi.org/10.1101/049288

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., … Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, *44*(D1), D710–D716. http://doi.org/10.1093/nar/gkv1157

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. P., & Wang, L. (2014). CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007. http://doi.org/10.1093/bioinformatics/btt730

APPENDIX A

Sample commands for various stages of the analysis pipeline.

- Preprocessing – Trimmomatic v. 0.36

Sample Command:

```
java -jar trimmomatic-0.36.jar \
         PE \
         -phred33 \
         -trimlog trimmomatic.log \
         read_1.fq  read_2.fq \
         read_1_trimmomatic.fq read_1_unpaired_trimmomatic.fq \
         read_2_trimmomatic.fq read_2_unpaired_trimmomatic.fq \
         ILLUMINACLIP: /adapters/TruSeq3-PE-2.fa:2:30:10 \
         SLIDINGWINDOW:5:20 \
         MINLEN:40
```

Parameter definitions:

PE – Paired End;

-phred33 – Data Encoded in the Phred-33 format (Illumina v.1.5);

-trimlog – Path to report run status;

ILLUMINACLIP – Path to the adapter library followed by three thresholds. The first value denotes the maximum number of mismatches of the seed and extend mapping procedure while the second and third determine the length of extension for paired-end and single-end reads, respectively;

SLIDINGWINDOW - Trim reads on a 5 bp window if the quality drops below 20 (from the 5' end);

MINLEN - Minimum read length.

- Read alignment - STAR 2.5.1b (Reference Index)

Sample Command:

```
STAR   --runThreadN 16 \
        --runMode genomeGenerate \
        --genomeDir data/ref/genome/index/STAR \
        --genomeFastaFiles data/ref/genome/GRCh38.primary_assembly.genome.fa \
```

```
        --sjdbGTFfile data/ref/annotation/gencode.v24.annotation.gtf \
        --sjdbOverhang 99
```

Parameter Definitions:

--runThreadN – Number of cpus to parallelize the job;

--runMode – Generate reference index;

--genomeDir – Path to output the index files;

--genomeFastaFiles – Path to the genome build Fasta file;

--sjdbGTFfile – Path to the file containing the splice junctions in .GTF format;

--sjdbOverhang – (Read length – 1).

- Read alignment - STAR 2.5.1b (Alignment)

Sample Command:

```
STAR   --runThreadN 16 \
        --runMode alignReads \
        --genomeDir data/ref/genome/index/STAR \
        --readFilesIn read_sample#_lane#_1.fq read_sample#_lane#_2.fq \
        --outFileNamePrefix read_sample#_lane#_star \
        --outSAMstrandField intronMotif \
        --outSAMtype BAM SortedByCoordinate \
        --twopassMode Basic \
        --outQSconversionAdd -31
```

- Linkage Disequilibrium

Sample Command:

```
plink   --bfile chr.dedup
        --r2
        --ld-snp-list chr_variants.txt
        --ld-window 100
        --ld-window-r2 0
        --out chr_variants_ld
```